Université de Montréal

The French Canadian founder population: lessons and insights for genetic epidemiological research

par
Héloïse Gauvin

Département de médecine sociale et préventive
École de Santé publique

Thèse présentée à l'École de Santé Publique
en vue de l'obtention du grade de Philosophiae Doctor (PhD)
en Santé Publique
option Épidémiologie

Août 2015

Université de Montréal
Faculté des études supérieures et postdoctorales

Cette thèse intitulée :

The French Canadian founder population: lessons and insights for genetic epidemiological research

présentée par :
Héloïse Gauvin

a été évaluée par un jury composé des personnes suivantes :

Dr. Anne-Marie Laberge, président-rapporteur
Dr. Philip Awadalla, directeur de recherche
Dr. Marie-Pierre Dubé, codirectrice de recherche
Dr. Marie-Élise Parent, membre de jury
Dr. Marc Tremblay, examinateur externe
Dr. Guillaume Lettre, représentant du doyen de la FES

Résumé

La population canadienne-française a une histoire démographique unique faisant d'elle une population d'intérêt pour l'épidémiologie et la génétique. Cette thèse vise à mettre en valeur les caractéristiques de la population québécoise qui peuvent être utilisées afin d'améliorer la conception et l'analyse d'études d'épidémiologie génétique. Dans un premier temps, nous profitons de la présence d'information généalogique détaillée concernant les Canadiens français pour estimer leur degré d'apparentement et le comparer au degré d'apparentement génétique. L'apparentement génétique calculé à partir du partage génétique identique par ascendance est corrélé à l'apparentement généalogique, ce qui démontre l'utilité de la détection des segments identiques par ascendance pour capturer l'apparentement complexe, impliquant entre autres de la consanguinité. Les conclusions de cette première étude pourront guider l'interprétation des résultats dans d'autres populations ne disposant pas d'information généalogique. Dans un deuxième temps, afin de tirer profit pleinement du potentiel des généalogies canadienne-françaises profondes, bien conservées et quasi complètes, nous présentons le package R GENLIB, développé pour étudier de grands ensembles de données généalogiques. Nous étudions également le partage identique par ascendance à l'aide de simulations et nous mettons en évidence le fait que la structure des populations régionales peut faciliter l'identification de fondateurs importants, qui auraient pu introduire des mutations pathologiques, ce qui ouvre la porte à la prévention et au dépistage de maladies héréditaires liées à certains fondateurs. Finalement, puisque nous savons que les Canadiens français ont accumulé des segments homozygotes, à cause de la présence de consanguinité lointaine, nous estimons la consanguinité chez les individus canadiens-français et nous étudions son impact sur plusieurs traits de santé. Nous montrons comment la dépression endogamique influence des traits complexes tels que la grandeur et des traits hématologiques. Nos résultats ne sont que quelques exemples de ce que nous pouvons apprendre de la population canadienne-française. Ils nous aideront à mieux comprendre les caractéristiques des autres populations de même qu'ils pourront aider la recherche en épidémiologie génétique au sein de la population canadienne-française.

Mots-clés : Population canadienne-française, épidémiologie génétique, partage identique par ascendance, analyse généalogique, génétique des populations, consanguinité lointaine, dépression endogamique.

Abstract

The French Canadian founder population has a demographic history that makes it an important population for epidemiology and genetics. This work aims to explain what features can be used to improve the design and analysis of genetic epidemiological studies in the Quebec population. First we take advantage of the presence of extended genealogical records among French Canadians to estimate relatedness from those records and compare it to the genetic kinship. The kinship based on identical-by-descent sharing correlates well with the genealogical kinship, further demonstrating the usefulness of genomic identical-by-descent detection to capture complex relatedness involving inbreeding and our findings can guide the interpretation of results in other population without genealogical data. Second to optimally exploit the full potential of these well preserved, exhaustive and detailed French Canadian genealogical data we present the GENLIB R package developed to study large genealogies. We also investigate identical-by-descent sharing with simulations and highlight the fact that regional population structure can facilitate the identification of notable founders that could have introduced disease mutations, opening the door to prevention and screening of founder-related diseases. Third, knowing that French Canadians have accumulated segments of homozygous genotypes, as a result of inbreeding due to distant ancestors, we estimate the inbreeding in French Canadian individuals and investigate its impact on multiple health traits. We show how inbreeding depression influences complex traits such as height and blood-related traits. Those results are a few examples of what we can learn from the French Canadian population and will help to gain insight on other populations' characteristics as well as help the genetic epidemiological research within the French Canadian population.

Keywords : French Canadian population, genetic epidemiology, identical-by-descent sharing, genealogical analysis, population genetics, distant consanguinity, inbreeding depression.

Table of Contents

List of Tables

List of Figures

Abbreviations and Acronyms

ACA :	Acadian population
AD :	Autosomal dominant
ADNmt :	Mitochondrial inheritance
AFR :	Africa
AME :	America
ANRIL :	Antisense non-coding RNA in the 9p21 locus
AR :	Autosomal recessive
BMI :	Body mass index
BP :	Blood pressure
BRCA1 :	Breast cancer gene 1 (or 2)
CaG :	CARTaGENE
CEPH :	Centre d'Étude du Polymorphisme Humain
CEU :	HapMap population from Utah residents with Northern and Western European ancestry
cM :	centiMorgan
CSASIA :	Central South Asia
CVD :	Cardiovascular diseases
CYP2C9 :	Cytochrome P450 2C9 gene
dbSNP :	National Center for Biotechnology SNP database
DNA :	Deoxyribonucleic acid
DNK :	Do not know
EASIA :	East Asia
EHR :	Electronic health record
EURO :	Europe
FC :	French Canadian
FEV1 :	Forced expiratory volume in one second
F_{ROH} :	Proportion of genome covered by runs of homozygosity
FRQS :	Quebec Health Research Fund
Fs :	Sum of LCAs' inbreeding coefficients
FVC :	Forced vital capacity
GFC :	Gaspesian French Canadian population
GWAS :	Genome-wide association study
HBA1c :	Glycated haemoglobin
HBD :	Homozygous-by-descent
HDL :	High-density lipoprotein
hg19 :	Human genome version 19
HHH :	Hyperornithinemia-hyperammonemia-homocitrullinuria
HLA :	Human leukocyte antigen
HMM :	Hidden Markov model
HSAN2 :	Hereditary sensory and autonomic neuropathy type 2
HWE :	Hardy-Weinberg equilibrium
IBD :	Identical-by-descent
IBS :	Identical-by-state

ID : Identification number

IMPQ : *Infrastructure intégrée des microdonnées historiques de la population québécoise*

kb : Kilobase

LCA : Lowest common ancestor

LD : Linkage disequilibrium

LDL : Low-density lipoprotein

LOY : Loyalist population

MAF : Minor allele frequency

MB : Megabase

MCH : Mean corpuscular haemoglobin

MCHC : Mean corpuscular haemoglobin concentration

MCMC : Markov Chain Monte Carlo

MCV : Mean corpuscular volume

MEDNIK : Mental retardation, enteropathy, deafness, peripheral neuropathy, ichthyosis, and keratoderma

MENA : Middle East and North Africa

Mix : Admixed individuals from different continental populations

MON : Montreal population

MRCA : Most recent common ancestor

NCBI : National Center for Biotechnology Information

NGS : Next-generation sequencing

NS : North Shore population

OMIM : Online Mendelian Inheritance in Man

PC : Principal component

PCA : Principal component analysis

PCA-CP : Principal component analysis on coancestry (chunk count) matrix

PCAgeno : Principal component analysis on genotypic data

PCSK9 : Proprotein convertase subtilisin/kexin type 9 (gene)

PQ : Whole sample from the Province of Quebec.

QUE : Quebec City area population

RCDW : Red cell distribution width

RMGA : *Réseau de Médecine Génétique Appliquée*

RNA : Ribonucleic acid

RNAseq : RNA sequencing

ROH : Run of homozygosity

SAG : Saguenay-Lac-St-Jean population

SD : Standard deviation

SNP : Single nucleotide polymorphism

VKORC1 : Vitamin K epoxide reductase complex subunit 1 gene

WBC : White blood cells

Xd : X-linked dominant

À mes parents, sans qui je n'aurais pu jouir de la vie

Acknowledgments

I would first like to thank Marie-Hélène Roy-Gagnon, my initial supervisor, for agreeing to take me as her PhD student. I felt from the very beginning that she trusted me fully, and introduced me to a number of interesting areas of research in genetics. Even after she left, she continued to provide advice and guidance.

On equal footing, I want to thank my advisor, Philip Awadalla, for letting me subsequently join his awesome lab. He offered me a wonderful opportunity. By his side, I have learned immensely about genetics and science as well as what makes a good scientist.

It was a great pleasure to work with the amazing members of the Awadalla lab, both past and present. Thanks to Vanessa Bruat, Jean-Christophe Grenier, Alan Hodgkinson, Armande Ang Houle, Julie Hussin, Youssef Idaghdour, Jean-Philippe Goulet, Jacklyn Quinlan, Marie-Julie Favé, Thibault de Malliard, Elias Gbeha, Mélanie Capredon, Mawussé Agbessi and Élodie Hip-Ki. Thank you for the lunches we shared, the encouragements you liberally handed out, the technical assistance you always offered and the general we had.

During my PhD years, I was also fortunate enough to meet Damian Labuda. Thanks for all the discussions and help. I also worked with people from his lab. Claudia Moreau and Jean-François Lefebvre, it was great to meet you along the way and I enjoyed working with you. I would also like to express my gratitude to Marie-Pierre Dubé, my co-supervisor, who provided encouragement, support and guidance.

I would also like to thank the Centre de recherche de Sainte-Justine. The staff is very welcoming to all students and offers valuable support. Special thanks to Sandy Lalonde, Dominika Kozubska, Alida Hounyovi and Cristina Pulciani.

My PhD would also not have been possible without the incredible BALSAC population database. Therefore I want to thank all the people who work there, especially Hélène Vézina and all those who set up this resource. Furthermore, I need to thank all participants who got

involved in the CARTaGENE Project. They are involved in a biobank which will contribute I am sure to great scientific discoveries: it has already started to. Thank you also to all the CARTaGENE team.

Moreover, I would like to express my gratitude to the administrative staff at the Public Health School: Monique Lespérance and Angélique DeChatigny. Thank you so much!

I am very thankful to all those who took time to comment my work or give insights on writing in English. Thanks also once more to the people who reviewed my papers.

I am deeply grateful to the members of my PhD jury for taking time out of their busy schedules to review this thesis.

During my doctoral studies, my research was supported by a variety of funding bodies. I want to thank the Fondation du CHU Sainte-Justine and the Fondation des Étoiles for their scholarships, as well as the Réseau de médecine génétique appliquée (RMGA) for the Louis-Dallaire fellowship. I also gratefully acknowledge the support I received from the Faculté des études supérieures et postdoctorales and from the Département de médecine sociale et préventive at Université de Montréal.

A special thanks to all the ladies who have been an inspiration to me: Marie Léger-St-Jean, Sarah Landry, Jacklyn Quinlan, Julie Hussin, Marie-Pierre Dallaire, Marie-Hélène Roy-Gagnon and super maxi-mom!

Enfin, je tiens à remercier mes amis, ma famille et belle-famille pour leurs encouragements. Merci!

Merci maman pour tout le soutien pendant mes années d'étudiante, je t'aime.

Et finalement je veux remercier Jean-Philippe pour son soutien et ses encouragements, mais surtout de me faire sourire chaque jour que je passe à ses côtés.

Chapter 1: Introduction

1.1. Genetic epidemiology

1.1.1. Historical perspectives

A hundred and fifty years ago Mendel established his principles of heredity, which dictate how genetic material is passed on. A century later, innovative molecular techniques revealed much more on genes and their structure and shed light on the high amount of variation existing between each individual's unique DNA sequence. Meanwhile the field of epidemiology, which began with the study of infectious diseases, slowly broadened its area of research to environmental factors, such as nutrition and air pollution. Eventually epidemiology developed up to a certain point where it became clear that genetic factors are important in the etiology and biology of diseases showing some degree of familial aggregation (Lilienfeld 1961). Environmental factors may also explain a part of the aggregation and need to be considered together with genetic factors.

Let's take as an example the Framingham heart study (Dawber *et al.* 1951). This long-term and still on-going study takes place in the town of Framingham, Massachusetts and was set up in 1948 to investigate cardiovascular diseases (CVD). This vast project contributed to the growth of epidemiology itself and now with its third generation of participants, Framingham study scientists are expanding their research to include the role of genetic factors (Jaquish 2007).

The alliance and interaction of genetics and epidemiology resulted in a new discipline called genetic epidemiology. This discipline is commonly defined as the study of genetic factors, their interaction and joint action with environmental factors, which are all leading to different distributions of disease in human populations, with the ultimate goal of control and disease prevention (Khoury *et al.* 1993; Thomas 2004). Genetic epidemiology initially started with recommendations for epidemiological studies to investigate the potential influence of genetic factors in the etiology of a disease (Neel and Schull 1954). This was quickly followed by the first discoveries of genetically defined groups (e.g. blood groups (Clarke *et al.* 1956))

associated to specific diseases. Genetic epidemiologists then outlined a basic process to study the genetic determinants of diseases. We will review the steps involved in this traditional research process in the next two paragraphs.

The traditional research process in genetic epidemiology begins with the description of disease risks varying among populations leading to the assumption that the disease is either related to the environment or to genetics or both. By looking more closely at the disease distribution it can be assessed whether familial aggregation occurs, meaning that the disease tends to occur more within families than in unrelated individuals. If familial aggregation is found, the next step is the segregation analysis, which is an approach used to study families with affected members to identify the most likely pattern of inheritance for the disease under study.

After the identification of the transmission model, individuals of interest are genotyped for known genetic markers and used in linkage analysis to attempt to pinpoint markers, which are almost all the time inherited by individuals having the disease, i.e. markers linked to the disease, and not or rarely inherited by disease-free individuals (Dawn Teare and Barrett 2005). In the end this step narrows down the search and finds roughly the disease gene location. Here the notion of recombination rate concerning the distance between markers is important. During meiosis, loci located closer together on a chromosome are more likely to be transmitted together. The more distant two loci are, the more likely it is that recombination will occur and separate them. Therefore the recombination rate between genetic markers along a chromosome and a disease gene provides information on its location. The following step is to fine map the disease gene since areas potentially containing disease genes identified through linkage are large. For this step we can take advantage of linkage disequilibrium (LD), which is a tendency for some alleles at linked loci to be associated with each other more than expected by chance, creating haplotypes (Palmer and Cardon 2005). Haplotypes are combinations of alleles at multiple linked loci that are transmitted together as they are found on a single chromosome. Affected individuals sharing overlapping segments of haplotypes can help localize the disease gene. Ultimately, sequencing a region to identify variants present in cases and absent in controls might be necessary to uncover causal variants or at least to describe the variation observed. Lastly a characterization of the gene involved in the disease is

done with regards to disease risk of the various mutation and possible interaction with age, sex or environmental factors.

The last two decades have seen many technological advances, which led to additional research tools and a broadening of the traditional genetic epidemiological research steps. Indeed, today's studies are not only relying on heritability within families and are now frequently carried out in cohorts of seemingly unrelated individuals having a disease, including relatively more complex diseases (polygenic) as opposed to Mendelian or monogenic diseases. Recent technological developments have broadened the scope of research approaches from family-based to population-based approaches such as genome-wide association studies (GWASes), which consist of scanning the whole genome looking for any variation associated with a disease or trait.

1.1.2. Advantages and limitations of genome-wide association studies

The sequencing of the human genome (International Human Genome Sequencing Consortium 2004), the HapMap project (The International HapMap Consortium 2005), which is a great catalogue of human haplotypes from around the world, and the advent of different databases listing genetic variants, such as the National Center for Biotechnology SNP database (dbSNP) (Sherry *et al.* 2001), allowed the realization of large-scale genotyping studies involving a large number of individuals.

GWAS is a popular large-scale population study approach, which has multiple advantages. Unlike candidate gene studies, a GWAS scans the whole genome with no a priori hypothesis regarding potential genes of interest and this unbiased approach offers a greater potential to make truly novel discoveries (Donnelly 2008). For example, new pathways were identified with GWASes providing further knowledge on mechanisms involved in the etiology of diseases and having direct clinical relevance (Visscher *et al.* 2012). GWASes rely on meiotic recombination events that have occurred in the history of a sampled population, as opposed to

meiotic events assessed in family studies by linkage analysis. As such, a region with a significant signal of association will be smaller at a genomic scale as compared to those found by linkage, which can facilitate the detection of the actual causal gene and mutation. Also, GWASes enable the discovery of genes with small effect sizes on disease risk, which are more difficult to track in linkage studies and are by definition more likely to contribute to common disorders than to familial ones. Indeed according to the now famous hypothesis that led to the popularity of GWASes, the "common disease, common variation" hypothesis, common allelic variation would account for a significant proportion of genetic variance in common disease susceptibility (Lander 1996; Schork *et al.* 2009).

The success of GWASes depends on the power to detect those associations between genetic variants and traits. Many factors influence this power, including the frequency of the risk genotype, the increase in disease risk associated with this risk genotype (relative-risk), the strength of the relation between the marker tested and the actual disease genotype, the sample size, the disease prevalence, the genetic heterogeneity of the sampled population, the accuracy of the genotyping technology used and the accuracy of the phenotype definition (or in the case of a quantitative trait its appropriate measurement) (Hattersley and McCarthy 2005; McCarthy *et al.* 2008). Very large sample sizes can be required to reach a reasonable power. This has favoured the formation of international collaborative consortia to aid recruitment of participants (Hattersley and McCarthy 2005).

In addition to the importance of having sufficient power, we have to consider a few limitations of GWASes. The most important one is probably that GWASes, as well as other genetic association studies, detect associations, which are not necessarily causal relationships (Cordell and Clayton 2005). As mentioned above, the large study sample size required can also be a limitation, especially if recruiting participants with a well-measured or homogeneous phenotype is difficult. Similarly the need for replication of results in independent samples in different populations, to increase evidence that the association is not an artefact due to uncontrolled variables, is also a major limitation. GWASes are interesting since they can detect associations with genetic variants having small effect but as a trade-off it typically means that their cumulative effect only explains a small fraction of an individual's risk for the

trait. Additionally, despite their large size, GWASes are typically less powered for rare variants (frequency < 5 %). Finally, GWASes identify location rather than gene and sometimes the variants found associated with a trait are far from coding regions or found in genes not thought to be related to the trait. However, this last drawback can also be worthy since new genetic variation and biological mechanisms can be disclosed.

In general GWASes will use genetic information from single nucleotide polymorphisms (SNP) arrays. This technology involves arrays of SNPs, which assay the most frequent form of variations in the genome. SNPs may or may not have functional consequences. Initially, arrays could type a few thousand SNPs and now, more recent ones type over 2 million variants (see (LaFramboise 2009) for more details). Generating such a large amount of information for every individual in large cohorts yields large amount of data to manage, which comes with some potential difficulties. Properly storing, managing and processing all this information have become increasingly challenging thereby parallel technological advances have addressed these challenges with the development of infrastructure and analysis pipelines.

1.1.3. Past successes and future directions

Genetic association studies led to an overwhelming number of discoveries. For example, within the framework of GWASes, over 15 000 SNPs were found associated to a disease (Welter *et al.* 2014). In the next section we review some of the most noteworthy discoveries.

For CVD the most famous genetic discovery is still the association of chromosome 9p21 to coronary artery disease and myocardial infarction (Samani and Schunkert 2008). Multiple studies reported at the same time an association for the same locus on chromosome 9 and in addition to being replicated in different ethnic groups, the effect of the locus was unaffected by traditional cardiovascular risk factors. The particular interest in this locus also comes from the fact that it is mostly deprived of coding genes and located in a region, which is well known in cancer genetics (Cunnington *et al.* 2010). The closest protein coding genes are about 100 kilobases (kb) away from the most strongly associated SNP and it was found that in fact the region of association is overlapped by a non-protein-coding RNA gene, called *ANRIL*

(Pasmant *et al.* 2007). Disease associated variants of the 9p21 region are highly correlated with *ANRIL* expression and this may suggest a modulation role in disease susceptibility (Cunnington *et al.* 2010). However it still remains unclear how 9p21 influences cardiovascular risks (Patel *et al.* 2014; Hannou *et al.* 2015).

Also related to CVD, the *PCSK9* gene has an important physiological role in cholesterol metabolism (Lambert *et al.* 2012). Different mutations on the gene were found to be associated with low-density lipoprotein (LDL) cholesterol levels. Both loss-of-function or gain-of-function mutations modify the availability of LDL receptor, which degrades LDL in a cyclic process and have a great impact on the levels of LDL cholesterol (Abifadel *et al.* 2003; Cohen *et al.* 2005, 2006). Since CVD are the leading cause of death in the world (World Health Organization 2015) and since *PCSK9* is recognized as a major factor influencing cardiovascular health, there is a growing amount of studies dedicated to potentials therapies targeting *PCSK9* (Awan *et al.* 2014; Dadu and Ballantyne 2014; Weinreich and Frishman 2014).

Genetic epidemiology also contributed considerably to advances in cancer research. Identification of breast cancer genes, through linkage analysis of families with cases of early-onset breast cancer (Hall *et al.* 1990) and confirmation that mutations on those genes were implicated in development of breast and ovarian cancer (Miki *et al.* 1994), opened the door to extensive efforts to characterize the genetic component of this cancer.

Along with the decreasing price of genotyping, the size of GWASes has correspondingly increased and the trend went on to creating larger and larger consortia. To facilitate research work, in the 2000s, a number of population-based biobanks were implemented and some established epidemiological studies redefined their mandates to include genetic aspects of diseases (Swede *et al.* 2007). Biobanks with vast sources of phenotypic information in addition to lifestyle, environment and other exposures assessment are great resources to perform research and to eventually translate genetic discoveries into clinical practice. They are also beneficial to epidemiological research in general, as data collection is often prospective and data collected allows studying the combined effects of different factors. As examples, the

most important population-based projects for now include UK Biobank with over 500 000 individuals recruited (Allen *et al.* 2012), Iceland's deCODE program with more than half of the 300 000 inhabitants population recruited (deCODE Genetics Inc. 2015), China Kadoorie Biobank with over half a million participants (Chen *et al.* 2011) and CARTaGENE project in the Quebec province with 40 000 participants (Awadalla *et al.* 2013; CARTaGENE 2015). Further details on CARTaGENE are presented in the section 1.3.4. (p.31).

Research approaches were also modified with the advent of next generation sequencing (NGS), which is producing tons of sequencing reads concurrently since the sequencing process was improved to be parallelized. Sequencing reads provide information on various types of genetic variation and SNPs called from them are largely free of ascertainment bias, thus providing a better picture of rare variations. Structural genetic variation such as copy-number variation is one type of variation drawing important research attention as it is purported to modulate gene expression and disease phenotype (Weischenfeldt *et al.* 2013).

1.2. Founder populations

1.2.1. Their advantages

Founder populations are populations that descend from a small number of founders, who left one area to settle in another and were isolated for some reasons, which can be linguistic, political, religious, social or geographic (Diamond and Rotter 1987). There has long been interest in these populations for genetic studies (MCKUSICK *et al.* 1964; Nevanlinna 1972; Skre 1975; Morgan *et al.* 1980) since they have numerous advantages compared to fully outbred populations with substantial admixture.

First, since the number of founders is limited when a new population is created, the amount of genetic variation carried is also limited and represents only a fraction of the genetic variation present in the entire population where the founders originated (see Figure 1.1). The founder effect and the population bottleneck are terms frequently used interchangeably however they

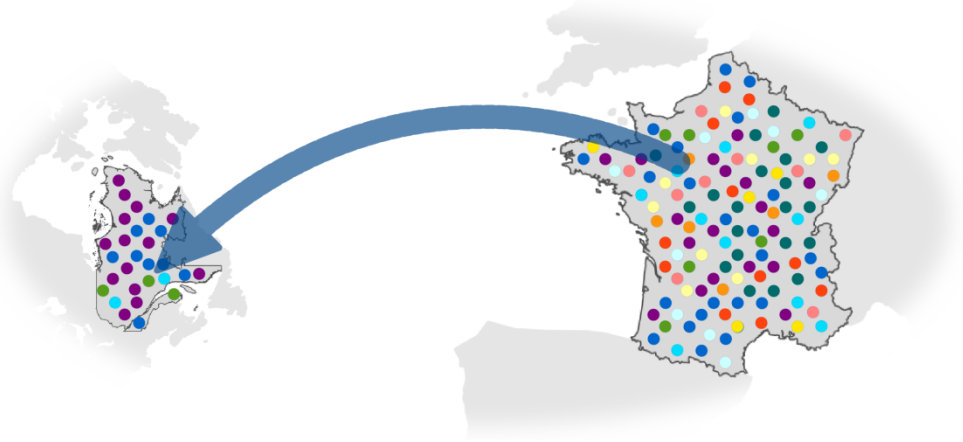

Figure 1.1 Founder effect

A non-random sample of the original population composed of multiple genetic members is isolated from the rest, thus a new gene pool with reduced genetic variation as compared to the original population is created.

are two slightly different concepts. The population bottleneck is a quick and sharp reduction in the size of a population due to a drastic event, which can also lead to a loss of genetic variation, while the founder effect concept puts emphasis on the loss of genetic variation caused by the creation of a distinct population. Founders will contribute only with their own genetic makeup to their offspring and so on for the next generations. Few new variants will appear, but the low mutation rate in humans (Conrad *et al.* 2011), and the isolation, or limited migration, will tend to reduce genetic variability. In some cases, this means that for a single disease or phenotype that may be caused by multiple different alleles, only some of these will be found in the founders (Shifman and Darvasi 2001; McClellan and King 2010). For example in the case of breast and ovarian cancer, over 2 000 variants in the genes *BRCA1* and *BRCA2* have been identified worldwide but only about 20 variants segregate among French Canadians (FC) and 3 mutations account for most of the FC breast cancers (Cavallone *et al.* 2010; Petrucelli *et al.* 2013). The potential reduced genetic heterogeneity also translates into a higher power to detect genetic variants. In other words it is easier to detect variants linked to a disease if instead of having dozens of those variants in a population you have only a few variants. This is true for simple Mendelian diseases but we can assume founder populations also have smaller sets of risk alleles involved in common diseases (Wright *et al.* 1999).

Another advantage of isolated founder populations stems from the fact that they often have experienced a rapid expansion, driving rare alleles to higher frequencies. Genetic drift, which is the change in the allelic frequency in a population due to random sampling at each generation, will act predominantly when the population is small (Masel 2011). Rare variants will not be the only variants influenced. Highly frequent variants can also become fixed, meaning every genome will have the same version of it, and some variants may also be lost. Of course, the increased frequency of some rare variants explains the increased prevalence of a number of recessive disorders in isolated founder populations, in which case, their discovery is facilitated (Puffenberger 2003). Note that recessive inheritance means that the mutated gene has to be transmitted by both the mother and the father in order for the child to have two copies of the mutation and to be affected by the disease. Recruitment of cases can also be facilitated when the disease prevalence is higher. Discovery of genetic factors is even more facilitated if the population also tends to have large kindreds, such as the Amish population,

which will also increase the odds of recruiting individuals with the exact same disorder (MCKUSICK *et al.* 1964).

Another advantage of founder populations is the relative homogeneous environment. Differences in diet, life habits, infectious agents, and sanitary conditions can be minimized for some isolated populations, therefore reducing environmental noise in a research framework. Since some isolated population also tend to be more geographically stable, i.e. not prone to move outside the community, this can also facilitate longitudinal and offspring studies.

Small population size also tends to translate into more inbreeding, which happens when related individuals are mating. More inbreeding means that recessive genes will be more likely to occur in a homozygous pattern, resulting in more recessive diseases and an opportunity to map these genes more efficiently. Also inbreeding translates into fewer generations separating randomly selected individuals (average time to the most recent common ancestor is smaller) and thus there are less recombination events between them (Shifman and Darvasi 2001). The lower recombination leads to greater correlation between segregating variants, i.e. the genetic units that are polymorphic within the population. Typically this means that more distant variants will tend to be inherited more often together, increasing the linkage disequilibrium (LD) between those variants (Shifman and Darvasi 2001). Patterns of LD depend on where recombination occurs and how genetic material is transmitted. Increased LD is advantageous to map disease variants. Indeed, when a region is well characterized, i.e. how its variants are linked to each other, and has a disease variant in high LD with other variants, these variants can be targeted instead of the disease variant itself (de la Chapelle and Wright 1998). In essence, this corresponds to a compromise between the extremely large segregating segments typical of linkage studies, and the very small shared segments typical of population-based GWASes.

Beyond LD blocks, the increased degree of relatedness found among individuals within isolated populations leads to the development and identification of large chromosomal segments. Patients afflicted by a disease in an isolated population will tend to share common ancestral haplotypes around the causal disease mutation, which can improve our ability to

detect such genes. Haplotype mapping can also give clues on the founder origin and on the time a mutation was introduced (Kibar *et al.* 2000; Yotova *et al.* 2005; Vézina *et al.* 2005a). When those haplotypes are inherited from a common ancestor (and consequently the same), they are called identical-by-descent (IBD) segments. Haplotypes and IBD segments are closely related concepts (see section 1.4.2 on p.35 for more details). Similarly runs of homozygosity (ROH) refer to regions of the genome where both chromosomal segments, inherited from the mother and the father, are identical. This may happen in situations where parents are more closely related. Homozygosity can also be used to map disease markers, especially in the case of recessive disorder (Lander and Botstein 1987; Bernard *et al.* 2010). The more inbred or closed a population, the higher the frequency of those long homozygous stretches, which will also contribute to facilitate genetic research in isolated populations.

Genetic epidemiological research can be more challenging when a lot of the data is missing or when the density of the genotypes is too sparse. Again, the genetic features of isolated populations (increased genetic homogeneity, increased LD and more shared tracts) can help with the imputation of genotypes and hence minimize the missing data problem (Marchini *et al.* 2007). Genetic imputation is a technique allowing one to infer unobserved genotypes (Li *et al.* 2009). Statistical imputation is also useful when merging data coming from different genotyping platforms that did not target the same variants.

One last major advantage of founder or isolated populations that deserves mention is that good genealogical record keeping is often available. There is substantial value to having reliable demographic history along with well curated extended genealogical data for the conduct of genetic research. The best known examples are Mormons with the Utah Population Database (Skolnick 1980), Icelanders with the Icelandic Genealogy Database (Tulinius 2011) and French Canadians with the BALSAC population database (Bouchard *et al.* 1989; BALSAC 2014a). These rich data, which are complementary to genetic information, represent an outstanding opportunity for research in several fields (Laberge 1969; Laberge *et al.* 2005a; Moreau *et al.* 2011). Information from genealogies can, for example, provide input on origin of genetic material, help to interpret modern population structure and to estimate time of gene flow events (Larmuseau *et al.* 2013a).

1.2.2. Challenges

There are numerous benefits that stem from the study of isolated population, and this has led to such populations attracting substantial interest from the genetic community (Peltonen *et al.* 2000). In the last decade, there has been renewed interest for isolated populations in part for rare variants studies recently possible through sequencing and also because of increasing computing resources allowing to build and to analyze large datasets, including large genealogical datasets (Holm *et al.* 2011). However, studying isolated founder populations is not challenge-free. First, reaching these populations is not always easy. In some instances, educational and linguistic barriers are present, or distrust of science can be encountered (MCKUSICK *et al.* 1964). In the specific case of Amish, the concern raised by the frequency of different rare disease led a doctor, Holmes Morton, to found in 1989 a non profit clinic devoted to the care, treatment and investigation of the genetic disorders afflicting the community of Lancaster County (Rosenblatt 2013). This kind of involvement in the community contributed to improving the education of Amish people about their specificity and raised their awareness on various health issues.

Also, not all populations have genealogical records of suitable quality and size amenable to research purposes. Input errors, when the information is digitized, may arise but quality controls and good management practices are usually set up especially when funded organizations oversee these operations (Bouchard *et al.* 1989; Cannon Albright 2008). However if the primary information is incorrect, errors cannot be avoided. For example, non-paternity or non-maternity events are one possible scenario, which can lead to mistakes. Assuming that non-maternity events are probably much less frequent than non-paternity events frequency estimation focused mainly on the last ones. Estimates of non-paternity events typically vary around 1-2% (Weir *et al.* 2006; Strassmann *et al.* 2012; Larmuseau *et al.* 2013b).

People from founder populations tend to be related to each other in many different ways, sometimes through inbreeding loops, which may complicate traditional segregation and linkage analyses. Alternative methods need to be used to disentangle family relationships.

Depending on genetic information availability, tracking of IBD segments can be an option to link individuals among relatives. IBD segments are a straightforward way to focus on the actual realized genetic sharing, as opposed to sharing expected from genealogical links, between individuals known to be related. Relatedness among sampled individuals can also induce bias in conventional association studies assuming independence and specialized tests are needed to account for known and unknown relationships between individuals (Yu *et al.* 2006; Sillanpää 2011).

As mentioned in the previous section, genetic drift is more likely to influence the distribution of allele frequencies in smaller populations, resulting in genetic variations specific to some populations and an increasing populations differentiation (Casals *et al.* 2013). An increase in the frequency of certain genetic variants may facilitate the detection of their association with a disease, but the scope of discovery and the genetic effect size may be reduced. In addition, the identification of disease variants specific to an isolated population may not be directly generalizable to other populations.

Another issue raised by the study of isolated populations relates to the selection of variants typically used on genotyping SNP arrays. Common genotyping platforms include SNPs that were discovered using sequencing technologies in different samples. These discovery panels are thus issued from genetic variations found in different populations of different sizes and the selected SNPs included in the platforms are not selected at random and may not necessarily be geographically representative (Wakeley *et al.* 2001; Clark *et al.* 2005). As a result, genotyping chips are typically biased towards common variants. This phenomenon is called ascertainment bias and can distort, among others, measures of population differentiation (Albrechtsen *et al.* 2010b). This means that, to some extent, common genotyping platforms may not be the best suited platforms to quantify and describe the genetic diversity in all populations. However alternative techniques such as sequencing can overcome this problem or correction techniques can be applied to infer measures of population differentiation (Albrechtsen *et al.* 2010b).

Finally ethical and social concerns need to be considered when working with isolated populations. How and what information is disclosed in publications is important for example

to avoid risks of collective stigmatisation, which can create damages that could eventually impair future participation to research studies (Bouchard 2004; Lavery *et al.* 2007). However, as stated earlier, educating participants about the potential risks, usefulness and benefits of their involvement as well as educating researchers about good communication practices offers a simple way to address this issue (Mascalzoni *et al.* 2010). There are also confidentiality and privacy issues that could be more difficult to manage in particular when genealogy data is linked to health records (Lavery *et al.* 2007; Mascalzoni *et al.* 2010). Note that most of these concerns and solutions also apply to non-isolated populations.

1.2.3. Examples

Some founder populations are better known than others as they have been more extensively studied. Table 1.1 (p.16) presents a wide number of founder, isolated and genetically differentiated populations. Comparing these populations help to understand differences among founder populations regarding 1) the genetic consequences of a bottleneck, 2) the demographic history, and 3) the cultural and environment features (Peltonen *et al.* 2000). Each population listed has a unique genetic background and contributed to our understanding of recessives diseases.

Note that other populations, with suspected founder effect or isolation, were not included in this list, either because documentation is lacking or the population did experience a considerable degree of admixture over time (the Cuban population for example (Cruz 2013)). As such, the table is not intended to present an exhaustive list (for more examples see also (Arcos-Burgos and Muenke 2002; Rudan 2006; Venken and Del-Favero 2007; Kristiansson *et al.* 2008)).

Table 1.1 Population isolates

Population name, location	History	Time of settlement	Isolation	Initial number of founders	Today's population size	References
Afrikaners, South Africa	Immigrants of Dutch origin settled in the Cape and later have spread inland.	1652	Religious and linguistic	~ 1 000	~3 000 000	(Diamond and Rotter 1987; Jenkins 1990)
Amish, North-East USA	Anabaptists from Switzerland settled initially in 3 counties, no admixture, closed population, high fertility rate.	1714-1727	Religious	~ 200	~290 000	(MCKUSICK *et al.* 1964)
Ashkenazi Jewish, Various locations	Jews moved from Alps to Rhineland ~9th century, moved to Eastern Europe ~12th century and then to Americas and Israel during 19-20th centuries. Had variable endogamy throughout history.	Not applicable	Religious and linguistic	Unknown	~10 000 000	(Ostrer 2001)
Finnish, Finland	Two early migration waves. Long term isolation. In the 16th century, migration from south and west coastal areas to north and eastern part of the country started creating sub-isolates. End of 17th century, famine and epidemics occurred followed by a rapid population size expansion.	~ 4 000 years ago	Mostly geographical	Probably small number	~5 000 000	(Peltonen *et al.* 1999; Kere 2001)
French Canadians, Quebec, Canada	French immigration until 1759, migration within the Quebec province creating several founder events and rapid expansion of the population	1608	Religious and linguistic	~ 8 500	~6 400 000	See section 1.3.1, p.18
Hutterites, Northern USA and Western Canada	Founders settled in 3 endogamous colonies, which have maintained separate identities and have high fertility rate	1875	Distinct geographic clusters, religious barriers	~ 440	~30 000	(Hostetler 1985; Abney *et al.* 2002)

Population name, location	History	Time of settlement	Isolation	Initial number of founders	Today's population size	References
Icelanders, Iceland	Few immigration, multiple bottlenecks	~ 900	Insularity	A few thousand Norwegian Vikings	~320 000	(Gulcher and Stefansson 1998; Tulinius 2011)
Mennonite North-East USA	Anabaptists from Switzerland settled in USA and population size increased over 150 years. From the end of the 18th century, multiple religious schisms led to splits of the population and bottlenecks.	1707-1757	Religious	3 000 – 5 000 immigrated to America	~390 000	(Puffenberger 2003)
Mormons, Utah, USA	Immigrants having mostly a British or Scandinavian ancestry settled in Salt Lake City and colonized the area. They have very high fertility rate.	~ 1847	Religious	~30 000	~1 750 000	(Jorde 1982; Slattery and Kerber 1993)
Newfoundlander Newfoundland, Canada	English Protestant and Roman Catholic Irish settlers, few new immigrants leading to homogeneous sub-populations.	~ 1610	Insularity, coastal outports	~ 20 000 (in 1760)	~525 000	(Rahman *et al.* 2003)
Sardinians, Sardinia, Italy	Low population density until 1700, many sub-populations (microgeographic heterogeneity)	~ 10 000 years ago	Insularity, geographical position, mountainous area	unknown	~1 600 000	(Calò *et al.* 2008)

Note: Some populations labelled under the same religious name are separate groups that do not necessarily share common ancestry (like Dutch-German and Swiss-German Mennonite (Orton *et al.* 2008)).

1.3. French Canadian founder population

1.3.1. A brief history of the peopling of Quebec

The study of the Quebec population is interesting because of its history, thus this section summarizes how *La Belle Province* was colonized. However, note that this is only a brief summary with a particular focus on elements and regions that are key to understand the history of FC participants in the Quebec Reference Panel and CARTaGENE, the two cohorts used in the context of this thesis.

The first French settlers founded Quebec City in 1608. Peopling was slow at the beginning essentially driven by the needs of fur trade (Charbonneau *et al.* 2000). Two other permanent settlements, *Trois-Rivières* and Montreal, were established respectively in 1634 and 1642 (see Figure 1.2, p.19). French immigration was coming from different parts of France but mainly the northwest departments and Paris region (Vézina *et al.* 2005b). This immigration has led about 8 000 settlers to leave descendants in the colony (Charbonneau *et al.* 2000). In fact, during the following period population increased mainly through reproduction. Specifically in the 18th century, the growth rate was so high that the population was doubling about every 30 years (Charbonneau *et al.* 2000). This growth rate was not even affected by the British Conquest in 1759 (Charbonneau *et al.* 2000). At that time the population was spread along both sides of the St. Lawrence River and comprised over 70 000 inhabitants (Charbonneau *et al.* 2000).

East of Quebec City, the peopling of the *Charlevoix* region started in 1675 and took place in coastal lowlands (Jetté *et al.* 1991). The region was relatively isolated due to surrounding mountains. The vast majority of the *Charlevoix* founders transited through another region of the colony before settling there and more than half of them were relatives (Jetté *et al.* 1991). At the beginning of the 19th century, the small region underwent some demographic pressure, migration to estuary heights began, towards inland, and overall there was an emigration flow

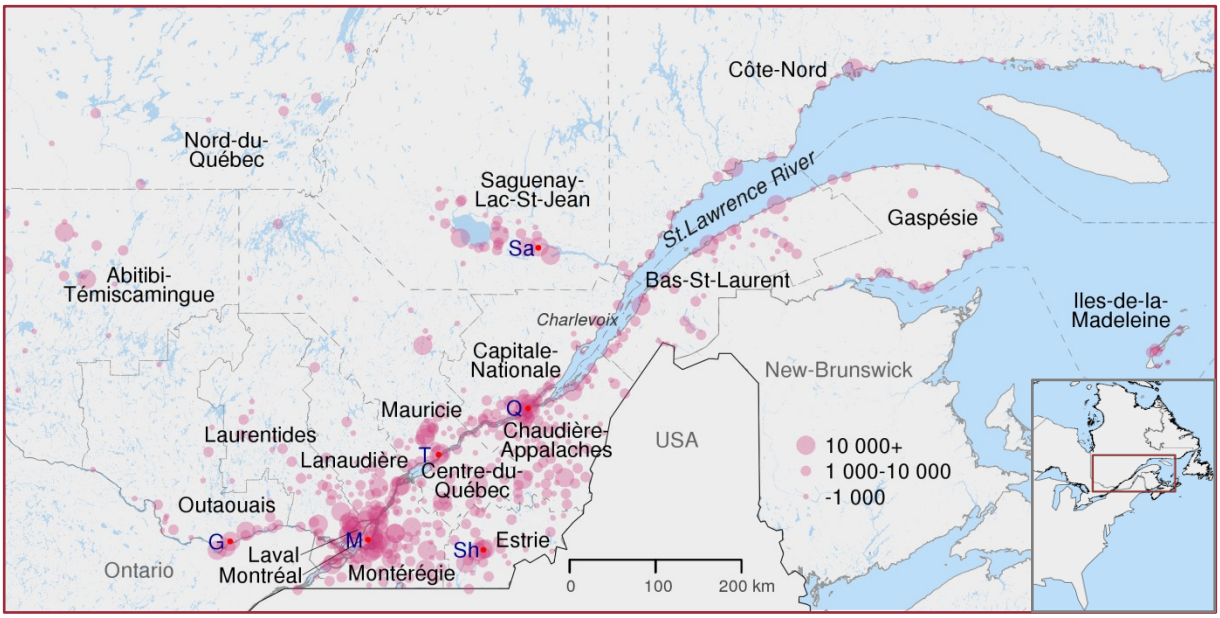

Figure 1.2 Map of the regions of Québec

Inhabited cities and villages are represented by pink dots sized according to the number of inhabitants in 2015. Dashed lines show present regions' boundaries. Red dots indicates some important cities : *G* Gatineau; *M* Montreal; *Q* Quebec City; *Sa* Saguenay; *Sh* Sherbooke; *T* Trois-Rivières.

out of the region leading to the settlement of *Saguenay-Lac-St-Jean* (referred as Saguenay in the following text) which started in 1838. Moreover, immigration to Saguenay was often carried out by members of the same families (Gauvreau *et al.* 1991). The rooting of these families in the region, their high fertility and their common origin contributed to shape the Saguenay population (Bouchard *et al.* 1988). After 1875, immigration to Saguenay was very low but the population size increased from about 5 000 in 1850, to 50 000 around 1900 and 250 000 in 1960 (Bouchard *et al.* 1988). The population size increase was due to natural growth; the birth rate was over 50 births per 1 000 people until 1930.

The North-Coast (*Côte-Nord*) region further north and east to Saguenay was under the same monopoly of the Hudson Bay Company for fur trade, as the Saguenay region, before a permanent coastline settlement began. Around 1830 we count some fur trading and fishing posts and less than a hundred persons living in the North-Coast all year long (Frenette 1996). Ending the monopoly led to an inflow of settlers interested in resource exploitation. People coming from *Iles-de-la-Madeleine* and Gaspesia settled in *Minganie*, the region including and facing Anticosti Island. Other parts of the region were populated with the establishment of sawmills and flour mills, the agricultural expansion, the development of the fisheries and mining (Frenette 1996). Incomers were from different parts of Quebec. The settlement of this large region was rather slow, in 1871 there is about 4 200 inhabitants excluding First Nations people. During the next 70 years, the population size had been multiplied by more than five (Frenette 1996).

All the way east, south of the St. Lawrence River, the Gaspé Peninsula has been frequented since the 16th century by seasonal fisherman mainly from Brittany, Normandy and the Basque country (Desjardins *et al.* 1999). French pioneers tried to establish the first permanent settlements in the mid-seventeenth century however, the efforts were hampered by many obstacles, such as underfunding, climate and war. It is only slightly before the British Conquest that settlers landed in permanently. After the Conquest, Gaspesia became a coveted territory. Fisheries as well as timber trade and shipbuilding industry finally developed. In addition to French, Gaspesia greeted Acadians, which were escaping deportation by the British (Bergeron *et al.* 2008). Acadians were also of French-descent but they did settle in 1604 in

another colony called Acadia (present-day Nova-Scotia, New-Brunswick and Prince-Edward-Island). A third important group, called Loyalists, also moved in Gaspesia. They were Anglo-American that wanted to remain loyal to the British Crown after the independence of the thirteen American colonies (Desjardins *et al.* 1999). These three groups tended to marry among themselves, which perpetuated until now their differences in terms of identity and genetic background (Vézina *et al.* 2014). Note also that Acadians populated other regions of Quebec and their integration with the local population varied. Gaspesia's population size went from about 3 000 in 1800 up to 50 000 a century later and peaked to 105 000 in 1961 before some decline (Desjardins *et al.* 1999).

Trois-Rivières was founded in 1634, west of Quebec City. About 30 years later, the settlement extended in the area partly due to the fur trade and in great proximity with the local First Nations (Hardy and Séguin 2008). The growth of the population of *Mauricie* was slow, counting about 4 000 inhabitants in 1760 (Hardy and Séguin 2008). Iron industry, farming and fur trading had been the initial drivers of development for the region and the commercial exploitation of forests started at the beginning of the 19th century. The strongest population growth in the region took place during the first half of the 20th century (Hardy and Séguin 2008).

Valued for its position as a gateway to the continent, the Montreal region eventually overshadowed the development of *Trois-Rivières*. Nonetheless, the beginnings of Montreal were slow and difficult; Iroquois, an important First Nations group, saw the settlers' arrival in 1642 as a threat. With its hospital, its fort and a population of over 500 settlers in 1660, Montreal was finally gaining importance (Linteau 2007). Early in Montreal's history, religious institutions occupied a large role, due among other things to the initial idea of a missionary settlement. In 1665, a French military intervention brought peace to the people of Montreal and opened the way for the expansion of fur trade. Montreal became the core of a commercial and political empire. The population grew less rapidly than in the St. Lawrence valley but was still over 4 000 in 1754 (Linteau 2007). After the British Conquest, English, Americans and Scottish immigrants took commercial control of the territory. In the 19th century, immigration led English-speakers to become the majority in the city. Trade was growing and agriculture all

around Montreal was developing. The second half of the 19th century was dominated by the industrialization of the city. In 1852 there were almost 60 000 inhabitants (Fougères 2012). British immigration eventually slowed, rural exodus began and the city expanded. Around 1866, French Canadians were again the majority of the population and by the end of the 19th century they did make up 60% of the population (>325 000 in 1925) (Fougères 2012). Montreal was home to two distinct populations (one of French-descent and the other of British-descent), each having their own institutions, their own schools, churches, etc. In the 20th century, immigration from all over the world contributed to population growth in Montreal and by the end of the century, the city grew to about 3 million inhabitants, including suburban areas.

West of Montreal, the St. Lawrence River continues south towards the Great Lakes. Another river, the Ottawa River also called the travelers' route, is located slightly up north and flows westward towards inland. Before 1800 *Outaouais* was a place of passage for the fur trade and was barely inhabited. After the Conquest, settlement started and the management of land dedicated to agriculture was established under the British system of townships (Blanchette 2009). Lands were used to attract Loyalists, English, Scottish and other immigrants. In 1800, the city of Hull (now part of Gatineau) was founded by an American, who came with his family. The region benefited from the rise of the lumber industry and the construction of the Rideau Canal, which started in 1820 on the other shore of the river in the future Canadian capital. In 1827 over 2 500 persons populated the region (Blanchette 2009). French Canadians were living alongside several other ethnic groups. Irish Catholics, among others, ensured that, with French Canadians, Catholicism became the main religion after 1840. In 1861 there were over 40 000 inhabitants and this number doubled 30 years later (Blanchette 2009). The late 19th century was characterized by the development of pulp and paper industry and by the time of the World War II, *Outaouais* was one of the most industrialized regions of the province (Blanchette 2009).

Throughout the 20th century, as in many industrialized countries, the rural exodus towards cities changed the demographic landscape of Quebec. In 1901, 40% of people were living in urban areas and it went up to 80% in 2001 (Piché and Le Bourdais 2003). Since 1965, the

population growth has decelerated mainly due to the drop of birth rate (Charbonneau 1973). The 20th century also saw the rise in immigration, which is increasingly diversified (Piché and Le Bourdais 2003).

Today the population of the Quebec province is mostly French speaking (overall close to 80%), except for some urban centers and the North populated by First Nations. Despite the British arrival with the Conquest in 1759, different factors prevented French Canadians to mingle with incomers. Language and religion were the most obvious factors of isolation. The Catholic composition of the French Canadian population favoured the Catholic Irish establishment over the Protestant branch (Grace 2003). Because they share the same religion Irish and French Canadians did mingle, resulting in the fact that about 20% of today's French Canadian population has an Irish ancestor (Tremblay *et al.* 2008). However, the Irish contribution to the French Canadian genetic make-up remains very small, about less than 1% (Tremblay *et al.* 2008). In the same way, the genetic contribution from other immigrant groups remained weak (Vézina *et al.* 2005b; Bherer *et al.* 2011). A study based on genealogical information reports that about 89% of Quebec's gene pool comes from French founders and almost 81% is derived from the French founders who arrived in the 17th century (Vézina *et al.* 2005b). French Canadians contributed to the growth of another population; between 600 000 and 800 000 French Canadians left Quebec for USA in 1840-1930 (Charbonneau 1973).

Note that we described the settlement of French people in *Nouvelle France* paying little attention to Aboriginal peoples who had already established themselves about 8 000 years ago in many regions (Frenette 1996; Desjardins *et al.* 1999). Although their influence has also shaped French Canadian history their genetic contribution to the population remained very small according to the latest studies (around 1%) (Vézina *et al.* 2012; Moreau *et al.* 2013).

The demographic history of Quebec during its early days is well known, thanks to the strong Catholic presence during colonization. The clergy and local authorities collected information about baptisms, marriages, burials in the Catholic population and also for converted individuals, e.g. some First Nations people (Charbonneau *et al.* 2000). In some cases, registers were active even before the erection of its parish (Hardy and Séguin 2008). All this

information was very well kept and well preserved and the wealth of nominal sources, such as notarized certificates and censuses, allowed to fill in the gaps and to confirm some information (Légaré 1988). Thereby a first population database combining many of those sources of information was set up by the Research Program in Historical Demography from *Université de Montréal* to cover all the population born in the St. Lawrence Valley under the French Regime (Légaré 1988; PRDH-IGD 2015). The Early Quebec Population Register includes over 2 400 000 records of vital events (birth, marriages and death certificates) (Desjardins 1998; PRDH-IGD 2015). Another population database, called BALSAC, started a few years later, in the 70s, with the computerization of records for the Saguenay region (Bouchard and De Braekeleer 1991). Researchers worked to pair 660 000 baptism, marriage and burial records for the whole period after 1842, a work that was later extended to cover the entire population of the province of Quebec with marriage certificates (Bouchard and De Braekeleer 1991). Today the BALSAC population database includes a total of 3 million records, including all catholic marriage certificates from 1621 until 1965, related to nearly 5 million people and genealogical reconstruction is still ongoing (BALSAC 2014b). A new project is also on the way as both population databases are working on the integration and matching of the data under the aegis of the integrated infrastructure of historical microdata on the Quebecers population (IMPQ).

Such extended genealogies for a relatively geographically stable and semi-closed population are pretty rare and developing the analytical and computational tool to take advantage of this valuable information is essential. One use of these genealogical data, which has probably attracted the most attention, is the study of the distribution of rare disorders in Quebec. Understanding the demographic history of Quebec is essential to explain the distribution of diseases and the presence of population heterogeneity.

1.3.2. Genetic profile of the French Canadian population

The French Canadian population evolved over the last four centuries in a unique way that led to different genetic characteristics. This section will focus on two of those; the structure of today's population and the presence of some particular inherited diseases.

Using a genealogical cohort with people married before 1800, Gagnon and Heyer (2001) confirmed the presence of an east-west gradient regarding population homogeneity. They explained how migration patterns can modify the symmetry or non-symmetry of the founders' genetic contribution. Specifically, first settlers to a territory tended to propagate more their genes than subsequent immigrants (in a continuous migration scheme) but when all immigrants entered a new region at the same time, then a more uniform genetic contribution to the newly formed population can be observed (one large move). They also highlighted the fact that at the end of the 17th century, the French Canadian population was already formed by three distinguishable groups.

Overtime the population continued to differentiate and today, the Quebec population is formed by multiple groups each corresponding to a regional or ethno-cultural population (Bherer *et al.* 2011; Roy-Gagnon *et al.* 2011). A study found a striking fit between the analysis of genealogical and genetic data and these analyses highlighted a population structure consistent with many founder events (Roy-Gagnon *et al.* 2011). It was a known fact that almost every French Canadian is related up to a certain point (Vézina *et al.* 2005b), however this study showed that relatedness patterns across the sub-populations of Quebec vary and are consistent with the demographic history (Roy-Gagnon *et al.* 2011).

Successive founder events leading to a particular population structure were the ideal setting for a higher incidence of different rare diseases. While some diseases are simply more frequent in Quebec, others are quite specific and almost unknown in the world, e.g. spastic ataxia of Charlevoix-Saguenay and agenesis of the corpus callosum (see Table 1.2, p.26). Some diseases are also occurring only in specific parts of the province like Gaspesia, Beauce, Bas-St-Laurent, Lanaudière and Saguenay. As it can be observed in Table 1.2, prevalence

Table 1.2 Examples of inherited diseases found in the French Canadian population

OMIM number[1]	Disease name	Type	Prevalence in Quebec[2]	Worldwide prevalence[3]	References
218000	Agenesis of the corpus callosum with peripheral neuropathy [CT]	AR	1/2 117	< 1/1 000 000	(De Braekeleer *et al.* 1993a; Dupré *et al.* 2003)
606002	Ataxia-oculomotor apraxia 2	AR	-	1/400 000	(Bouchard *et al.* 1980; Duquette *et al.* 2005)
604370, 612555	Breast cancer (susceptibility to)	AD	-	1/400	(Tonin *et al.* 1998; Chappuis *et al.* 2001; Cavallone *et al.* 2010)
118220	Charcot-Marie-Tooth disease type 1A	AD	-	1/10 000	(Dupré *et al.* 1999)
129500	Clouston hidrotic ectodermal dysplasia	AD	-	1/10 000 – 1/100 000	(Clouston 1929; Kibar *et al.* 2000)
602579	Congenital disorder of glycosylation type Ib	AR	-	< 1/1 000 000	(Pelletier *et al.* 1986; Vuillaumier-Barrot *et al.* 2002)
219700	Cystic fibrosis	AR	1/902	1/8 000	(Daigneault *et al.* 1991)
219800	Cystinosis	AR	1/6 237	1/100 000 – 1/200 000	(Richler *et al.* 1991; McGowan-Jordan *et al.* 1999)
143890	Familial hypercholesterolemia	AD	1/81 –1/270	1/500	(De Braekeleer 1991; Vohl *et al.* 1997)
238600	Familial hyperchylomicronemia	AR	1/5 000 – 1/8 000	<1/1 000 000	(Gagne *et al.* 1989)
300624	Fragile X syndrome	Xd	-	1/4 000 – 1/6 250	(Rousseau *et al.* 1995; Dombrowski *et al.* 2002)
229300	Friedreich Ataxia	AR	-	1/50 000	(Bouchard *et al.* 1979; Keats *et al.* 1987b)
235200	Haemochromatosis	AR	-	1/1 000	(de Braekeleer *et al.* 1992; Rivard *et al.* 2000)
201300	HSAN2	AR	-	<1/1 000 000	(Roddier *et al.* 2005)
238970	HHH syndrome	AR	-	1/8 300	(Camacho *et al.* 1999; Debray *et al.* 2008)

OMIM number[1]	Disease name	Type	Prevalence in Quebec[2]	Worldwide prevalence[3]	References
614615, 614970	Joubert syndrome	AR	-	1/100 000	(Joubert *et al.* 1969; Srour *et al.* 2012b, a)
602390	Juvenile hemochromatosis	AR		<1/1 000 000	(Rivard *et al.* 2003)
220111	Lactic acidosis [CT]	AR	1/2 473	1/40 000	(De Braekeleer 1991; Morin *et al.* 1993)
535000	Leber's hereditary optic neuropathy	ADNmt	-	1/10 000 – 1/100 000	(Macmillan *et al.* 1998; Laberge *et al.* 2005a)
609313	MEDNIK syndrome	AR	-	< 1/1 000 000	(Saba *et al.* 2005; Montpetit *et al.* 2008)
252500	Mucolipidosis II	AR	1/6 184	< 1/1 000 000	(De Braekeleer 1991; Plante *et al.* 2008)
160900	Myotonic dystrophy	AD	1/630	1/7 000 – 1/50 000	(Yotova *et al.* 2005; Mathieu and Prévost 2012)
164300	Oculopharyngeal muscular dystrophy	AD	1/1 000 – 1/1 750	1/10 000 – 1/100 000	(Brais *et al.* 1995; Duquette and Giard 1997)
261600	Phenylketonuria [SP]	AR	-	1/20 000	(John *et al.* 1990; Carter *et al.* 1998)
264700	Pseudovitamin D-deficiency rickets	AR	1/2 916	1/2 000 – 1/10 000	(De Braekeleer 1991; Labuda *et al.* 1992)
610743	Recessive ataxia of Beauce	AR	-	< 1/1 000 000	(Dupré *et al.* 2007; Gros-Louis *et al.* 2007)
270550	Spastic ataxia of Charlevoix-Saguenay [CT]	AR	1/1 932	-	(De Braekeleer *et al.* 1993b; Engert *et al.* 2000)
609041	Spastic paraplegia 27	AR	-	< 1/1 000 000	(Meijer *et al.* 2004)
182601	Spastic paraplegia 4	AD	-	1/20 000 – 1/50 000	(Meijer *et al.* 2002, 2007)
600354	Spinal muscular atrophy	AR	-	1/33 000	(Simard *et al.* 1997)
272800	Tay-Sachs disease	AR	-	1/333 000	(Keats *et al.* 1987a; De Braekeleer *et al.* 1992)

OMIM number[1]	Disease name	Type	Prevalence in Quebec[2]	Worldwide prevalence[3]	References
276700	Tyrosinemia type 1 [CT, SP]	AR	1/1 845	< 1/1 000 000	(Laberge 1969; De Braekeleer and Larochelle 1990; Phaneuf *et al.* 1992)
302800	X-linked hereditary neuropathy	Xd	-	1/7 000	(Hahn *et al.* 1990; Dupré *et al.* 2001)
214100	Zellweger syndrome	AR	1/12 191	1/50 000	(Levesque *et al.* 2012)

[1]Online Mendelian Inheritance in Man (OMIM) is a comprehensive and well curated catalogue of human genes and genetic disorders accessible through http://omim.org (McKusick-Nathans Institute of Genetic Medicine 2015). [2]Prevalence for Quebec are most of the time specific to a subpopulation. [3]Some worldwide prevalence estimates are specific to Caucasian people or people with a European ancestry. [SP]Included in the Québec Newborn Blood and Urine Screening Program. [CT]Included in the carrier testing pilot-project for Saguenay region. Type of inheritance are: *AD* Autosomal dominant; *ADNmt* Mitochondrial inheritance, *AR* Autosomal recessive; *Xd* X-linked dominant. Abbreviations*:* *HSAN2* Hereditary sensory and autonomic neuropathy type 2; *HHH syndrome* Hyperornithinemia-hyperammonemia-homocitrullinuria syndrome; *MEDNIK syndrome* Mental retardation, enteropathy, deafness, peripheral neuropathy, ichthyosis, and keratoderma syndrome.

rates for a number of diseases are much higher in French Canadians. This fact have drawn attention towards the French Canadian population who was the object of several genetic reviews (De Braekeleer 1991; Scriver 2001; Laberge *et al.* 2005b; Dupré *et al.* 2006). Note also that those reviews contain other examples of diseases as Table 1.2 is not an exhaustive list.

Although most of the diseases found were recessive, inbreeding was early on discarded as a reason for the emergence of this disease burden. Inbreeding coefficients computed from genealogical data for a group of disease cases of Saguenay were compared to those of a group of non-affected individuals and were found to be not much higher (De Braekeleer and

Gauthier 1996). Researchers also observed very few marriages of second-degree or close cousins among both groups (De Braekeleer and Gauthier 1996). Another study compiled a more comprehensive picture for the whole province and showed that close consanguinity rates varied across Quebec sub-populations and were overall small while distant consanguinity, having ancestors related within 13 generations, reflects the situation of almost all French Canadians (Vézina *et al.* 2004).

The French Canadian population is a very good population for gene discovery and gene identification is facilitated by the founder effects, the presence of many cases and the potential higher homogeneity (of disease phenotypes and genotypes) (Engert *et al.* 2000; Srour *et al.* 2012b; Chetaille *et al.* 2014). Some of the latest gene discoveries have been done through FORGE (Finding of Rare Disease Genes), a Canadian consortium, that is a collaborative effort to study rare diseases by using next-generation sequencing technology (Beaulieu *et al.* 2014). Efforts towards a better understanding of these diseases were also driven by the interest to provide appropriate health services and care to the population. Hitherto different public health initiatives have been set up in the province to help to prevent, detect, investigate and cure those diseases. The Quebec Newborn Blood and Urine Screening Program includes some of those diseases (annotated with ^{SP} in Table 1.2), which can have severe consequences but that may be prevented if the disease is detected before its onset (Gouvernement du Québec 2015). Especially as some diagnoses are easier and cheaper to obtain through a biochemical or phenotypic analysis than a mutated genotype. Older initiatives, such as clinics dedicated to neuromuscular and metabolic diseases and targeted genetic counselling, also offer particular support to sick individuals and their family. More recently a pilot-project, which is still ongoing was implemented to provide carrier testing for all the population in Saguenay for four major recessive hereditary diseases (annotated with ^{CT} in Table 1.2) (Pouliot and Rousseau 2014). Information sessions are held and carrier testing is performed on a voluntary basis. The main goal of carrier-testing in this case is to give people information in order that they will be able to make informed reproductive decisions. Additional genetic counselling and psychosocial intervention services can be provided to couples that are both carrier of the same disease. Genetic counselling and prenatal diagnosis are actions that can lower the number of births at risk (Mathieu and Prévost 2012).

1.3.3. The creation of a biobank for Quebec

At the turn of the new millennium, the CARTaGENE project was being set up and a team of researchers was working to establish what was going to become the biggest biobank in the Quebec province. At that time, people were becoming familiar with new medical genetic concepts and the advent of community genetics was promoted. The initial aim of the CARTAGENE project was to set up the first genetic map of Quebec by recruiting 1% of its population (Lacroix 2001). This vast project was intended to uncover the genetic origin of different complex diseases, such as cancer and cardiovascular diseases, and to become a major tool for public health and genetic epidemiology (Lacroix 2001). CARTaGENE was developed with initial funding help from Réseau de Médecine Génétique Appliquée (RMGA) of the Quebec Health Research Fund (FRQS) and from Genome Quebec and the Health Ministry of the Quebec government. Later on Genome Canada and the Canadian Partnership Against Cancer Project, an organization which has the mandate to foster the fight against cancer and other chronic diseases for the benefit of all Canadians (Canadian Partnership Against Cancer Corporation 2015), also provided funding.

From the very beginning of the CARTaGENE project, the community participation was encouraged through workshops and consultations (Avard *et al.* 2009). A partnership approach was set up in order to involve the public in the decision-making processes (Godard *et al.* 2004). The approach established wanted to help to maintain public trust in biomedical research and to better understand community concerns about privacy, transparency, accountability, discrimination and stigmatization (Godard *et al.* 2004).

The first major phase of recruitment was held from July 2009 to October 2010 and over 20 000 participants were then part of the cohort (Awadalla *et al.* 2013). In 2012 two re-contact efforts were done to collect additional information on the environment and the nutrition (Awadalla *et al.* 2013). From the end of 2012 and during the two following years 20 000 more persons were recruited (CARTaGENE 2015).

1.3.4. CARTaGENE overview

In 2015 CARTaGENE is now a large population-based cohort encompassing a vast amount of information for more than 40 000 Quebecers from the greater Montreal, Quebec City, Saguenay-Lac-St-Jean, Sherbrooke, Trois-Rivières and Gatineau. Participants are aged between 40-69 years and have almost all visited one of the 13 assessment sites for data collection. Detailed questionnaires on health and socio-demographic status were captured, physiological measures were taken and biological sample (blood, serum and urine) were collected (Awadalla *et al.* 2013). All data is stored on highly secure servers and biological specimens are kept in a dedicated and secure high-tech biologic repository in Chicoutimi.

CARTaGENE uniqueness includes high depth phenotyping and a genealogical division. Indeed, the BALSAC population database joined CARTaGENE project early in its development. That alliance has led CARTaGENE to propose an additional questionnaire to participants in order to recover their genealogical information, a questionnaire that was filed by over 28% of the participants (CARTaGENE 2015).

Today, CARTaGENE is a resource that has been used by more than forty research efforts, some of which has already published their discoveries (CARTaGENE 2015). One important project using the initial set of 20 000 participants, has made the headlines when it was uncovered that many Quebecers are not aware that they are sick (Verhave *et al.* 2014; Daoust-Boisvert 2014). Another outstanding investigation has looked at mitochondrial RNA among ~700 CARTaGENE participants. Hodgkinson *et al.* (2014) ultra-deeply sequenced mitochondrial transcriptomes and found an impressive amount of variations, while previous studies reported that variation in these sequences was rare. But that is not all: they also found that post-transcriptional modification in mitochondria was largely driven by a missense mutation in a gene known as being involved in mitochondrial transfer RNAs processes. Another study having great repercussions showed, still using CARTaGENE project, that mutational burden within individuals was clearly modulated by recombination rates along the genome (Hussin *et al.* 2015). These varying patterns were observed in the French Canadian population as well as other populations with different demographic histories.

Ongoing projects are quite diverse and numerous, including, amongst others, studies on the metabolic syndrome, the emotional well-being and health status, the impact of range population expansion on the genome and the influence of genetics and the environment on cardio-metabolic phenotypes in Quebecers (CARTaGENE 2015).

1.4. Relatedness

1.4.1. Theoretical expectations

A recurring element in the studies of French Canadians is the level of relatedness they have. Thanks to the work of great pioneers in the world of quantitative genetics, that are Sewall Wright, Ronald Aylmer Fisher and Gustave Malécot, early on they have defined kinship coefficients to measure the degree of relatedness of a pair of relatives and the degree of inbreeding of a single individual (Fisher 1918; Wright 1922; Malécot 1948). We will present this basic theory here.

First remember each genome has half its genetic material from a mix of a father genome and the other half from a mix of a mother genome. The mixing, which is called genetic recombination, occurs during meiosis in order to create gametes. Briefly during the meiosis step chromosomes, which are coming in pairs, are duplicated and homologous chromosomes exchange genetic material. Then, chromosomes split twice to end up with reproductive cells (gametes) with a single copy of each chromosome. All our cells are said diploid, which means they have two sets of chromosomes, apart from gametes which are haploid (one set). Understanding how the transmission of genetic material occurs we are now able to figure out the principle of close relatedness and its implication on the genetic material shared by relatives. Obviously a parent-child will share exactly half of their genetic material. For all other scenarios of relatedness, the randomness of sexual reproduction intervenes. This means that for a grandparent-grandchild relationship we can expect that on average they share one fourth of their genome, but in reality it can be a little bit more or less. All theoretical expectations can be summarized with the use of kinship coefficients. Kinship coefficients are defined as the probability that two alleles randomly selected from each of the two individuals are identical-by-descent (IBD), i.e. are identical and coming from the same common ancestor.

The kinship between individuals *i* and *j* will be defined with this formula:

$$\varphi_{ij} = \sum_{a} (1 + f_a) \frac{1}{2}^{g_{ai}+g_{aj}+1}$$

which is the summation over all the lines of descent of ancestor *a*, where g_{ai} (g_{aj}) is the number of generations between subject *i* (*j*) and ancestor *a* and f_a is the inbreeding coefficient of ancestor *a*. Inbreeding coefficient for an individual is equals to its parents' kinship coefficient and is zero when parents are unrelated. When two individuals are linked through only one ancestor, which is not inbred, formula can be simplified to:

$$\varphi_{ij} = \frac{1}{2^{g+1}}$$

where *g* is the number of generations separating them. This calculation method is called the path-counting approach and gets quite complex as soon as the number of common ancestors increases. For the ease of computation most methods implement the recursive formula:

$$\varphi_{ij} = \frac{1}{2}(\varphi_{if} + \varphi_{im})$$

where *f* and *m* are the father and mother of *j* (Karigl 1981). The expected proportion of alleles that are IBD between two individuals is $2\varphi_{ij}$.

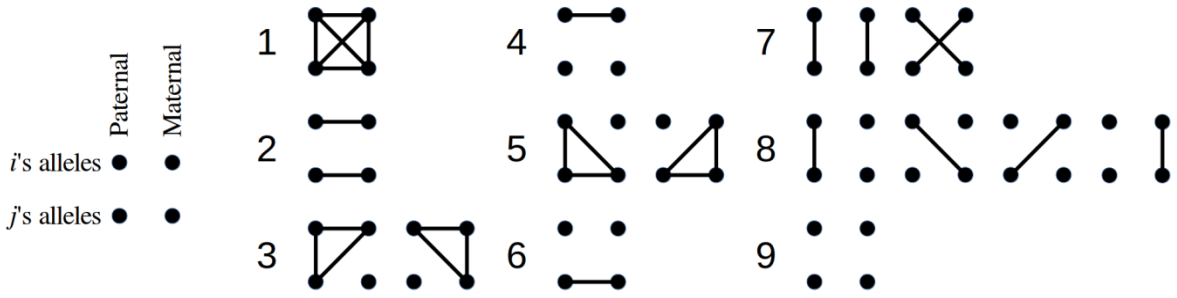

Figure 1.3 The fifteen identity states grouped in nine condensed states

Detailed identity states for individual *i* and *j*, nodes are alleles and lines indicate alleles that are IBD. Adapted from Jacquard 1974.

Jacquard refined the definition of those kinship coefficients with genetic identity coefficients, which discriminate all possible scenarios of genetic relation (Jacquard 1974). Figure 1.3 shows the 15 detailed identity states, which can be summarized in 9 condensed identity states. First 6 states assume one or both individuals are inbred. With those states we can compute kinship this way:

$$\varphi_{ij} = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8$$

where Δ_k is the probability of the condensed state *k* (see Figure 1.3). This alternative is not necessarily computationally more straightforward but helps to visualize every little detail underlying a relationship. In order to achieve the calculation the use of a computer with suitable tools is inevitable unless the relationship is fairly simple with very few common ancestors. Many programs allow those computations since it is a basic one for genealogical analysis (Lee *et al.* 2010; Lange *et al.* 2013).

1.4.2. Identical-by-descent and observed sharing

When genealogical information is available theoretical expectations will provide valuable information about genetic sharing. On the other hand, when no familial data is available to rely on, one option is to investigate genetic information. By interrogating the genetic material we can get an idea of the actual realized (as opposed to expected) genetic sharing between two individuals and we can use this information to speculate on their relationship. Different estimators exist and their use depends on the kind of genetic information available. Here we will focus on genetic variation coming from single nucleotide polymorphisms (SNP).

The simplest measure of relatedness between two individuals is the genetic similarity. For each genetic variant we wonder if it is the same or not in each person, i.e. is it identical-by-state (IBS) or not ?, and we compute the proportion of variants, which are IBS. Average IBS pairwise identity computed on variants distributed genome-wide is providing basic

information, which can help to discriminate homogeneous subsets like different populations, but remains limited for relationship inference.

To go a step further and to aim for higher degree relationships, we can ask whether genetic variants were inherited IBD or not. However the difference between IBS and IBD is quite thin and assumptions have to be done in order to state whether or not it is the same variant coming from a recent common ancestor (Powell *et al.* 2010). Actually one of the subtleties of the definition just got exposed; the track of time is very important. Depending of the time scale considered it can be more or less easy to look for common ancestors. For example, going back to a few thousand years we could all be cousins (Rohde *et al.* 2004). Although we could argue on the recentness of this time scale here we will consider a much smaller scale, which is fitting the IBD resolution and spans a couple of hundred years.

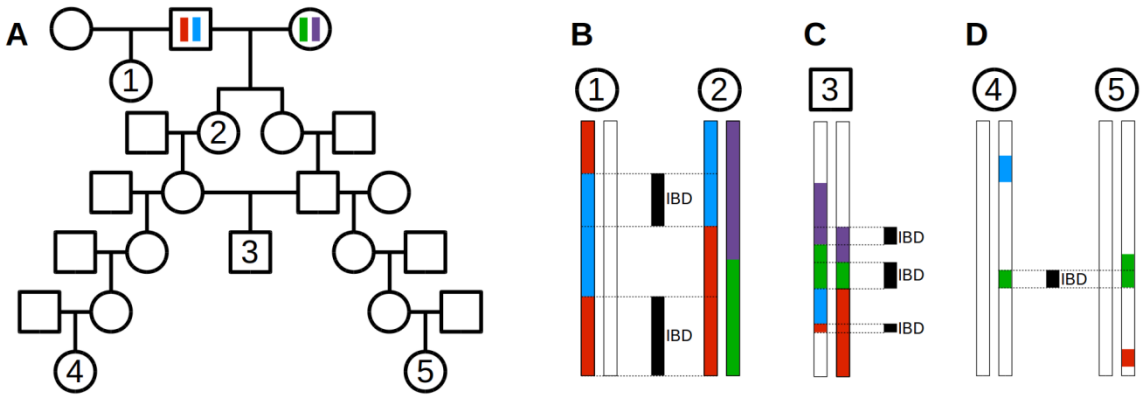

Figure 1.4 Identical-by-descent transmission

A) Six generations pedigree. Square symbols are used to represent males and circles are used for females. Colored bars are chromosomes; left bar is from the paternal lineage and right bar is from the maternal lineage. B) Two half-siblings with a father in common. Regions of IBD are shown with black bars. C) Offspring from a first-cousin relationship, black bars are homozygous by descent. D) Two fourth-cousins. Adapted from Heutink and Oostra (2002) and Browning and Browning (2012).

So since variants are transmitted through chromosomes, they are coming in stretches. In other words it means a genetic variant alone is weakly informative about its origin but flanking variants forming a segment are much more informative. Notions of haplotypes and LD, which were introduced in section 1.1.1, are particularly relevant here. Recombination breaks parental chromosomes to form remixed genetic material, so haplotypes are transmitted over the generations but they are eventually getting smaller and cut in pieces (see Figure 1.4). Therefore the ability to label variants as IBD or not relies on the ability to build those haplotypes.

When genetic information is acquired there is no track of which variants comes from the maternal and paternal chromosomes. Determining this particular information is an important step called phasing the data (Tewhey *et al.* 2011). Phasing can be achieved with different computational techniques depending on complementary information available (Browning and Browning 2011a). It is possible to phase one person's genetic information by using population-based approaches relying on statistical and computational techniques that capitalize on LD patterns. The advent of multiple population panels and a project dedicated to the reconstruction of haplotypes, HapMap, were very useful for this step and they surely contributed in making this technique the most used one. In a little bit more ideal scenario, the person's genome was analysed with other persons from the same population and they will all be used in a concurrent way to elucidate everybody's phasing information. The ideal option, which is also the less cost-effective, is to leverage information also from both parental lineages (through trio recruitment) to decipher which material was inherited from a person's mother and father. Other experimental techniques, which are still in development, not yet cost-effective on a large-scale, and to be further validated, involve the physical separation of homologous chromosomes or the use of very long sequencing reads (Browning and Browning 2011a). Recent statistical tools phase data mostly using a hidden Markov model (HMM) to find haplotype conditional on the individual's genotypes (Scheet and Stephens 2006; Browning and Browning 2007; Howie *et al.* 2009; Li *et al.* 2010; Delaneau *et al.* 2012). Phasing data also helps when it comes to genotype imputation (Marchini and Howie 2010). Major challenges of phasing involve the computation time, which varies with the dataset size and the SNPs density, and their accuracy, which can be measured with the switch-error rate

(Browning and Browning 2011a). Switch errors happen when the inferred haplotype phase needs to be switched a number of times to match the true haplotype phase, in other words when some haplotypic segments' connections are wrong.

The definition of haplotype phase opens the door to identify chromosomal parts that two individuals may share IBD. Multiple approaches have been implemented so far. Most IBD inference methods are either relying on a haplotypic dictionary, or using a HMM on phased data or performing phasing simultaneously to IBD inference (Gusev *et al.* 2008; Browning and Browning 2011b; Han and Abney 2011; Palin *et al.* 2011). Each method has its own specific parameters to enhance inference on particular populations in addition to common parameters to handle missing data, genotyping errors and the most important ones, parameters to define IBD regions.

Whether IBD inference will be a matter of a high certainty defined with a probability or by a certain minimal segment length or both depends on how IBD status is modeled within each method. Broadly as a shared segment increase in length we have an increased confidence that the segment is truly IBD. Indeed as the random process goes on and on, the size of segments transmitted from a generation to another decreases by a factor equivalent to the number of meioses (m) separating the two relatives, specifically segments will have an expected length of $100/m$ cM (Browning and Browning 2012). The more distant two relatives are the more difficult the IBD segments will be to tract as their sizes decrease to a point where it becomes too difficult to distinguish from LD blocks. Meanwhile the segment sizes decrease, the average proportion of the genome shared IBD will also decrease but much faster and will be $1/2^{m-1}$ (Browning and Browning 2012). Those simple theoretical expectations get much more complicated when the shared ancestry is not owed only to one but multiple different degree ancestors such as in a founder population. Moreover genome-wide patterns of IBD sharing have not been described for remote and complex relatedness as it can be observed in the French Canadian founder population. Therefore this particular point will be one of the topics studied in the second chapter of this thesis.

Besides the main interest of IBD, that is being informative for relationship inference, IBD segments have many purposes. IBD tracts can be used to investigate population history (Gusev *et al.* 2012; Palamara *et al.* 2012; Moreau *et al.* 2013; Gravel *et al.* 2013) and to map diseases (Houwen *et al.* 1994; Kenny *et al.* 2009). IBD mapping is somehow in between association studies, which are testing single variants, and linkage analysis in families, which is looking for variants cosegregating with the disease variants while it does not require any pedigree or genealogical information to be performed. IBD mapping can even be more powerful than single variant association when rare variants are involved (Browning and Thompson 2012) or more powerful than analysing families independently when those ones have a common ancestral background (Glazner and Thompson 2012). Since IBD detection is done pairwise the amount of information quickly rises and it can get tricky to identify groups of interest but methods to cluster IBD segments have been developed to overcome this issue (Gusev *et al.* 2011; Qian *et al.* 2014). IBD sharing can also be used to compute heritability and estimates coming from it may be even less biased than estimates coming from close relatives (Zuk *et al.* 2012). Distant relatives are less purported to share a common environment, which can bias heritability estimates. As for phasing, once IBD status is inferred it can be leverage to improve imputation accuracy (Kong *et al.* 2008).

To fully benefit from IBD segments a better understanding of its patterns and of the information, which can be extracted for relationship inference, is definitely important. Most IBD inference methods development and assessment relied on simulated datasets and a framework with real genotypic data and genealogical data could allow uphold their validity. Moreover few methods exist to infer relationships and they basically infer only simple relationships (Weir *et al.* 2006; Huff *et al.* 2011), i.e. individuals related through one or two common ancestors, so there is plenty of work to do to provide details on how related are two individuals regardless of the level of complexity from the characteristics of IBD segments.

1.4.3. The case of inbreeding

We presented relatedness between individuals and how we can group certain equivalent relationship states (Figure 1.3). There is a particular case we did not discuss a lot, which is the relatedness with oneself, i.e. inbreeding, and which corresponds to states 1 to 6 on Figure 1.3 (p.34). When an individual has parents who are somehow related this results in an offspring having parts of its homologous chromosomes coming from the same common ancestor (see Figure 1.4C, p.36). Inbreeding coefficient for individual a (f_a) is equals to φ_{fm}, where f and m are its father and mother and unless they are unrelated the value is greater than zero.

In terms of IBD sharing inbreeding is also called being homozygous-by-descent (HBB) since the two copies are coming from the same ancestor (autozygosity) necessarily they will be homozygous. Therefore the more inbred a person is, the more homozygous genetic material he has. Initially inbreeding was a concern for all elements of organisms' vigor (weight, fertility, vitality, etc) hence the term "hybrid vigor" is often used for the opposite case (Wright 1922). Wright was explaining in 1922 that mutations are more likely to cause damage than improve an organism qualities and damaging dominant mutations are more likely to be quickly withdrawn leaving the recessive ones to accumulate. He was absolutely right consequently when inbreeding occurs; homozygosity occurs for random genes and allows recessive traits to be expressed. This is why homozygosity itself was used early on to try to map diseases suspected to be recessive (Smith 1953; Lander and Botstein 1987).

Here our attention will not be focused on one disease; we will attempt to assess how inbreeding depression (opposite to hybrid vigor) can influence some health-related traits. Many studies reported that an increased homozygosity was associated with quantitative traits such as blood pressure, cholesterol, height, intelligence (Rudan *et al.* 2003b, a; Campbell *et al.* 2007; Woodley 2009; McQuillan *et al.* 2012; Panoutsopoulou *et al.* 2014), and was even a predictor of coronary heart disease, stroke, cancer, depression, asthma, gout, and peptic ulcer, (Rudan *et al.* 2003a). Association were also found with height, weight and BMI in children (Fareed and Afzal 2014). As for IBD sharing, to evaluate inbreeding we can use different indices such as the raw rate of homozygosity (or heterozygosity) or the length of the

homozygous stretches called runs of homozygosity (ROH), and different methods (Purcell *et al.* 2007; Han and Abney 2011; Gusev *et al.* 2012; Browning and Browning 2013a). However those methods as well as the availability of very dense genetic data are quite recent so most previous studies on inbreeding depression were using either reported information on consanguinity (inbred or not), either computing inbreeding coefficients on small pedigrees or using some microsatellites data to assess homozygosity. With the advent of SNPs we think we can get a better picture of the extent of the actual homozygosity and then assess whether or not it is associated with a range of physical, haematological and biochemical traits. Moreover the French Canadian population, which will be under study, is thought to have more distant than close inbreeding, which could reduce or eliminate its consequences.

Consanguinity is an important phenomenon with rates varying across populations and history and it is believe that 10% of today's world population is derived from union of second degree cousins or closer relatives (Bittles and Black 2010). In the French Canadian population, the inbreeding myth persists today and especially about the Saguenay region, where many recessive disorders have been reported. It has been shown however that marriages between close cousins were no more frequent in Saguenay than in any other region of Quebec (Vézina *et al.* 2004). In fact what turns out to be true is that Saguenay has one of the highest rates of distant consanguinity, which is calculated considering all ancestors over the last 13 generations (Vézina *et al.* 2004).

1.5. Research questions and thesis outline

In the previous sections we reviewed how genetic epidemiology has evolved and how founder populations contributed to knowledge about diseases, genetic burden and genetic in general. We presented the French Canadian population along with the current state of knowledge about its history, its population structure, its disease burden and its special toolkit (extensive genealogical data and an important biobank) available for the research community.

French Canadians are definitely a population from which we still have a lot to learn. With its load of genealogical information it is the perfect framework to study and compare the demographic mechanisms that shaped today's population. The way the peopling of the Quebec province occurred led to unique patterns of relatedness among its inhabitants. Thereby the main goal of this thesis is to understand how demographic history shaped the genetic sharing in French Canadians and how these genetic relationships can explain variation in different health traits.

Inferring relatedness from genetic patterns is interesting but what if we had extensive genealogical information? No such comparison of extensive genealogical data, which can be seen as expected sharing between individuals, and genome-wide dense genetic, which can be used to measure relatedness have ever been done to our knowledge. Moreover, the study of relatedness patterns can help to improve our knowledge of French Canadians and ease inference of relationships in other founder populations. Therefore in the second chapter we present a study investigating different methods to infer the regions of genetic sharing among a group of French Canadian individuals coming from different regions of the Quebec and describing this genetic sharing along with the genealogical characteristics.

Handling an impressive amount of genealogical information can be tricky and describing adequately the collected information in not an easy task either; unless we have the proper tool to do it. Most softwares and packages we know include particular sets of functions and could be more comprehensive and better suited for extensive genealogical data, which generally spans several generations and includes a large number of individuals. In the third chapter, we

present a tool developed in order to analyze extensive genealogical data. The R package called GENLIB includes basic descriptive functions in addition to functions to simulate gene-dropping in a genealogy. We provide a detailed example of how the tool can be used on a genealogical corpus of over 40 000 persons and perform a simulation study to investigate the population characteristics, which are influencing the genetic sharing among people.

Knowing that inbreeding depression can have a negative impact on health we wonder how important this phenomenon is, in particular when consanguinity is quantified with the latest techniques, and if we can detect it among French Canadians. The French Canadian population is known to have sub-populations with distant consanguinity but it is not clear if inbreeding over an interval of several generations also has a worthwhile impact on health phenotypes. Many studies of inbreeding depression used pedigree-based estimates to assess inbreeding and we question this approach. In the fourth chapter I present a study on the impact of distant inbreeding on multiple health-related traits. We use a sample of individuals from the biobank CARTaGENE and target individuals having a French Canadian origin to measure their level of inbreeding. We use their inbreeding coefficients to determine the presence of correlations with their physiological traits.

The last chapter summarises results from the three studies and outlines the limits of this work. I also suggest new research questions and I discuss the impact of this work in a broad context and in relation to knowledge of the Quebec population.

Chapter 2:
Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population

Héloïse Gauvin, Claudia Moreau, Jean-François Lefebvre, Catherine Laprise, Hélène Vézina, Damian Labuda and Marie-Hélène Roy-Gagnon

Reference: Gauvin H, Moreau C, Lefebvre J-F, Laprise C, Vézina H, Labuda D, Roy-Gagnon, M-H. Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. *European Journal of Human Genetics*, 2014; **22**: 814–21. doi:10.1038/ejhg.2013.227

Authors' contribution

In this paper, my contribution is:

- A literature review on IBD detection methods;
- The design of the study with MHRG, DL and HV;
- Statistical analyses;
- Writing of the paper.

Contributions of other authors are: CL, DL and MHRG contributed samples. CM and JFL provided bioinformatic support and performed experiments. CM, JFL and MHRG analysed the data. HV provided genealogical reconstructions. All authors reviewed and approved the manuscript.

Acknowledgements

I am grateful to all participants who generously shared their DNA and information required to reconstruct their genealogies. I also thank all the anonymous reviewers and editors for their valuable comments on the analyses, which helped improve this manuscript.

This work was supported by a grant from the Fonds de la Recherche en Santé du Québec (FRSQ) to MHRG. Support from the FRSQ Réseau de Médecine Génétique Appliquée (RMGA) and the Canadian Institutes of Health Research (CIHR; to DL and HV) is also gratefully acknowledged. HG is the recipient of a scholarship from the RMGA.

Abstract

In genetics the ability to accurately describe the familial relationships among a group of individuals can be very useful. Recent statistical tools succeeded in assessing the degree of relatedness up to 6-7 generations with good power using dense genome-wide SNP data to estimate the extent of identity-by-descent (IBD) sharing. It is therefore important to describe genome-wide patterns of IBD sharing for more remote and complex relatedness between individuals, such as that observed in a founder population like Quebec, Canada. Taking advantage of the extended genealogical records of the French Canadian founder population, we first compared different tools to identify regions of IBD in order to best describe genome-wide IBD sharing and its correlation with genealogical characteristics. Results showed that the extent of IBD sharing identified with FastIBD correlates best with relatedness measured using genealogical data. Total length of IBD sharing explained 85% of the genealogical kinship's variance. In addition, we observed significantly higher sharing in pairs of individuals with at least one inbred ancestor compared to those without any. Furthermore, patterns of IBD sharing and average sharing were different across regional populations, consistent with the settlement history of Quebec. Our results suggest that, as expected, the complex relatedness present in founder populations is reflected in patterns of IBD sharing. Using these patterns, it is thus possible to gain insight on the types of distant relationships in a sample from a founder population like Quebec.

Introduction

In genetics research, the ability to accurately describe the familial relationships among a group of individuals can be very useful. For example, genome wide association studies generally assume that studied subjects are independent and this assumption can be assessed easily if the list of their recent ancestors is known and error-free. Almost everybody can identify their parents and generally also their grandparents or even great-grandparents. However, most people do not know about their ancestors more remote than two or three generations unless extensive genealogical records are available for the population studied, as in the cases of the Hutterites (Abney *et al.* 2000), Icelanders (Tulinius 2011) or Amish (Khoury *et al.* 1987) for example.

Another way to describe relationships among individuals in a dataset is to look directly at their genome. Recent statistical tools succeeded in assessing the relatedness up to 6-7 degree relatives with good power using identity-by-descent (IBD) sharing (Huff *et al.* 2011). IBD sharing, estimated with genome-wide single nucleotide polymorphism (SNP) data, is defined as segments of the genome shared identically between two individuals. These chromosome segments are identical-by-state (IBS) and descend from a common ancestor without occurrence of any recombination event (Powell *et al.* 2010). A segment IBD is always IBS but the reverse is not necessarily true unless the timescale is unlimited. In practice, IBD detection from SNPs captures relatively recent ancestry since the resolution of IBD segment detection in a specific dataset limits the timescale that can be considered (Browning and Browning 2012).

Following the important technological innovations that made large amounts of genome-wide SNP data available at reasonable costs, several methods to detect IBD sharing between individuals have been developed. Approaches are generally based on the likelihood that a genetic sequence is IBD, which is measured with a probabilistic model detailing the whole IBD process or using the frequency of haplotypes, where low frequencies of a shared haplotype is an indication of highly probable IBD, or by setting a segment length threshold as a sequence is more likely to be IBD as it is spanning a large chromosomal segment. For example, GERMLINE is a method using a length threshold that builds up a dictionary with

chunks of haplotypes and IBD segments are spotted in accordance with a minimal length and with some flexibility as genotyping errors might be present (Gusev *et al.* 2008). The most flexible method is the hidden Markov model (HMM) that provides a basic framework to which probabilities for genotyping error and a linkage disequilibrium (LD) model can be added (Purcell *et al.* 2007; Albrechtsen *et al.* 2009; Browning and Browning 2011b; Han and Abney 2011). Haplotypes or genotypes can be used and some inference methods also use IBD detection to improve or to perform phasing (Kong *et al.* 2008; Genovese *et al.* 2010; Palin *et al.* 2011). Simulations studies have shown that more complex models had lower false discovery rates and higher sensitivity, in particular higher power to detect small segments, resulting in greater accuracy of IBD segment detection (Gusev *et al.* 2008; Browning and Browning 2011b; Han and Abney 2011; Palin *et al.* 2011).

Most comparisons of IBD inference methods have been conducted in homogeneous, unstructured populations and in simulation frameworks. In fact, to our knowledge, IBD inference methods have not been compared in a real-data setting with extensively documented genealogical records, and genome-wide patterns of IBD sharing have not been described for remote and complex relatedness such as that observed in the French Canadian founder population of the province of Quebec, Canada. The history of the French Canadian founder population begins with French settlers arriving at the beginning of 17th century (Charbonneau *et al.* 2000). Immigration from France ceased with the British Conquest in 1759. From 1755, Acadians, who were descendants of French pioneers who settled in Acadia (located in areas of present-day Nova Scotia, New Brunswick, and Prince-Edward Island), started to move to several regions of Quebec, escaping the deportation led by the British (Bergeron *et al.* 2008). In the last part of the 18th century, American Loyalists, who wanted to stay under the British rule, also moved to Quebec. Meanwhile, the French Canadian population expanded rapidly in relative isolation caused by linguistic, religious and geographic barriers, which amplified the founder effect (Bouchard and De Braekeleer 1990). As population size grew, settlers colonized new regions of Quebec, including remote and isolated regions, which resulted in population structure (Bherer *et al.* 2011; Roy-Gagnon *et al.* 2011).

In this study, we focused on three regions: the Saguenay-Lac-St-Jean, the western part of the North Shore and the Gaspe peninsula, as well as the two main cities of the province, Montreal and Quebec City (Fig. 1 in Roy-Gagnon *et al.* (2011)). In Saguenay, French Canadian settlement started around 1840 with the arrival of inhabitants from the neighbouring region of Charlevoix. Between 1840 and 1910, 75% of the 30,000 immigrants to Saguenay came from that region (Pouyez *et al.* 1983). The region of the North Shore was mainly colonized by people from the Charlevoix and Bas-St-Laurent regions between 1840 and 1920 (Frenette 1996). On the other side of the St. Laurence River, in Gaspesia, permanent European settlement began some decades earlier. In the second half of the 18th century the Gaspe Peninsula first greeted Acadians. Soon after, Loyalists joined them. Lastly, French Canadians attracted by developing fishing, naval and lumber industries also moved to Gaspesia (Desjardins *et al.* 1999). These three groups then evolved quite separately as they married mostly among themselves (Desjardins *et al.* 1999).

During the 19th and 20th century, immigration from various origins mixed into the French Canadian population with a very limited genetic impact and it has been shown that early founders have a greater contribution to the current gene pool (Heyer *et al.* 1997; Bherer *et al.* 2011). Today, about 80% of the 8 million inhabitants of the province is French speaking (Statistics Canada 2012).

The availability of genealogical data is a major advantage for genetic research in Quebec. Two important population registers exist: the BALSAC population register and the Early Quebec Population Register. The information contained in these databases comes primarily from vital statistics (births, marriages, deaths). As of November 2012, the BALSAC population register contained over 3 million records, which have been computerized and linked to cover the whole province for the 19th and 20th centuries (mostly marriage records) (BALSAC 2012). The Early Quebec Population Register contains all records from the beginning of settlement (1608) to 1800 for a total of 700,000 records (Desjardins 1998). Using these population registers, it is possible to reconstruct ascending genealogies of subjects from the present-day population going back over four centuries.

In this study, we used extensive genealogical data from Quebec in combination with genome-wide SNP data to first compare inference of IBD sharing provided by different methods in order to best describe genome-wide IBD sharing and its correlation with genealogical characteristics. IBD sharing detection was performed on a sample including seven populations of Quebec: French Canadians, Acadians, and Loyalists from Gaspesia as well as French Canadians from Saguenay-Lac-St-Jean, North Shore, Quebec City, and Montreal. Our analyses showed a good correlation between total length of IBD sharing and genealogical kinship coefficients for most methods with FastIBD yielding the best correlation overall. Using IBD results from FastIBD, we found differences in genome-wide IBD sharing patterns across sub-populations, which reflect genealogical characteristics. This information suggests that IBD sharing can reveal, at least in part, the complex relatedness present in a sample from a founder population like Quebec.

Material and Methods

Study population

The data consist of 143 individuals from a previously reported sample from seven sub-populations of Quebec (Roy-Gagnon *et al.* 2011). Recruitment criteria focused on the geographical origin of participants, and as much as possible, we recruited participants with at least one parent born in the region before 1960 or who were themselves born in the region before 1960. For all individuals, using the BALSAC population register and the early Quebec Population Register, genealogies were reconstructed as far back as possible and confirmed the absence of closely related individuals (first cousins and closer) in the sample. All participants gave their informed consent and the CHU Sainte-Justine Ethics Committee approved the study protocol.

For comparison purposes, we downloaded the original CEU sample (II+III) from the International HapMap project (International HapMap Consortium *et al.* 2007). We excluded 2 highly related individuals (Pemberton *et al.* 2010), leading to a set of 109 individuals (more distantly related than cousins) with North-Western European origin (see Supplementary Table 2.2 p.72).

Genotyping and quality control

Sample from Quebec was genotyped on Illumina HumanHap650Y arrays at the McGill University and Genome Quebec Innovation Center. Quality control procedures were the same as in the first publication using this sample (Roy-Gagnon *et al.* 2011). Briefly, quality check was performed to retain individuals and SNPs with at least 90% genotypes and to select only common autosomal SNPs (MAF > 5%) in Hardy-Weinberg equilibrium (exact test (Wigginton *et al.* 2005), $p>0.001$). These restrictions yielded 140 individuals (20 Gaspesian French Canadian, 20 Acadians, 20 Loyalists, 22 from Saguenay-Lac-St-Jean, 20 from the North Shore, 16 from Quebec City and 22 from Montreal) and 539 742 SNPs. The same quality

control criteria were applied to HapMap CEU, yielding 538 776 SNPs. All genomic positions are according to NCBI build 37.

Genealogical data and associated measures

The completeness of the genealogical data is measured by the proportion of ancestors observed (i.e. ancestors for whom information is available) in the data at a given generation divided by the expected number of ancestors. The completeness of our genealogical data of our sample of 140 individuals is over 90% up to the 5th generation and over 80% up to the 9th generation, except for the Gaspesian Loyalists (see Roy-Gagnon *et al.* (2011) for a more detailed description of completeness in these data). The lower amount of genealogical information available for the Loyalists sample is mainly due to their later arrival in Quebec and, to a lesser extent, to the fact that Protestant records were less complete and less well kept than Catholic records (which cover French Canadians and Acadians).

To describe the sample, kinship and inbreeding coefficients were calculated using the S-Plus® 8.0 (S-PLUS 8.0. Copyright 1988, 2007 Insightful Corp) function library GenLib. This library implements the algorithm of Karigl to calculate kinship coefficients (Karigl 1981). We also used PedHunter software (Lee *et al.* 2010) to get the set of lowest common ancestors (LCAs) for each pair of individuals. LCAs are the most recent ancestors shared by a pair of individuals. A pair can have more than one LCA as long as no ancestor in the set of LCAs shares a descendant who is also an ancestor of the pair of individuals. We also obtained, using PedHunter, the length of the shortest paths from one member of the pair to the other member through their LCAs, named hereafter distances to LCAs. Once each set of LCAs was obtained, we calculated the inbreeding coefficients of these LCAs. We used the sum of these inbreeding coefficients to measure the total amount of inbreeding present among the LCAs.

Genomic IBD sharing

We selected five different methods to perform the detection of IBD segments, all using a probabilistic framework except GERMLINE. GERMLINE is a computationally efficient software implementing a method that builds a dictionary of haplotypes to find matches between individuals. These matches are then extended to identify long shared segments while allowing some flexibility by assuming an error rate per SNP in order to avoid too many false negatives caused by genotyping inaccuracies (Gusev *et al.* 2008). Other methods are largely based on hidden Markov models (HMM). PLINK is the simplest method since it does not allow genotyping error and assumes that SNPs are in approximate linkage equilibrium (Purcell *et al.* 2007). IBDLD incorporates potential genotyping errors and missing data and has an extension for LD (Han and Abney 2011). The FastIBD method also includes a LD model when estimating IBD. The inference is conducted on sampled haplotypes for which an IBD score is calculated using shared haplotype frequency. Detected tracts are then extended and identified as being IBD according to a threshold set on score values (Browning and Browning 2011b). The last method that we considered, SLRP, also uses a HMM to approximate the IBD process while considering a genotyping error rate (Palin *et al.* 2011).

For all methods default parameters were used and some data manipulations were performed when necessary (Supplementary Table 2.3 p.73). For PLINK, which does not include LD, we did SNP pruning (pairwise $r^2 < 0.2$ in sliding windows of size 50 shifting every 5 SNPs) leading to a subset of 65 959 SNPs. For GERMLINE, we phased data with two different methods; Beagle version 3.3.1 (Browning and Browning 2007) and ShapeIT version 1.378 (Delaneau *et al.* 2012). For all analyses, we kept only segments greater than or equal to 2 cM, corresponding to the expected length of segments for common ancestors up to 25 generations ago (Browning and Browning 2012; Brown *et al.* 2012). This length ensures a good sensitivity and limits the false discovery rate (Gusev *et al.* 2008; Browning and Browning 2010, 2011b; Brown *et al.* 2012).

Statistical analysis

We first examined the correlation between IBD sharing identified with the different methods and genealogical kinship coefficients. We used the total length of all segments shared IBD and calculated Pearson's correlation coefficients. Assuming that genealogical kinship is the true expected kinship, we selected the method providing the best correlation as the best method for our population and retained this method for further analyses. We also examined the distribution of the lengths of the IBD segments identified by each method and we considered computation time.

We then examined the relationships between genomic IBD sharing and genealogical characteristics using simple linear regression models. We also looked at genomic sharing in pairs of individuals with or without at least one inbred LCA. Lastly, we investigated differences in IBD among the sub-populations. We plotted the average number of segments of a certain size shared per pair of individuals and also the proportion of pairs of individuals having IBD sharing at each position on the genome.

Results

Genealogical description

Levels of relatedness among individuals within the different sub-populations, as measured by the kinship coefficients estimated from the genealogical data, vary greatly (Supplementary Figure 2.5 p.69). As described in Roy-Gagnon *et al.* (2011), people from Saguenay and North Shore as well as Acadians had higher levels of kinship while populations from Montreal and Quebec City areas were less related. These observations are consistent with the settlement history of the province of Quebec and are also supported by previous findings based on genealogical data that emphasized a West-East decreasing gradient of diversity among regional populations as well as a stratification of regional populations (Bherer *et al.* 2011).

Figure 2.1A (p.57) presents the distributions of the number of LCAs per pair of individuals excluding unrelated pairs according to the genealogical kinship coefficients (i.e., pairs of individuals with kinship = 0). Pairs of individuals with kinship value equals to zero are pairs unrelated relatively to the time scale considered or related but without enough genealogical information available to support the relationship. In the whole sample, the average number of LCAs per pair of individuals was 74, ranging from 2 to 433. Average numbers of LCAs for the sub-populations ranged from 2.3 (Loyalists) to 152.6 (Montreal area), and the distributions were significantly different among sub-populations (all Kolmogorov-Smirnov test p-values < 0.007). For each related pair we also looked at the distance to the most recent LCA and the mean distance to LCAs (Figure 2.1C, D), which were on average 15.5 (ranging from 5 to 24) and 19.8 (ranging from 7 to 24), respectively. These distributions were also significantly different across populations (p-values < 0.02) except for minimal distance to LCA for Loyalists compared to Acadians and Gaspesian French Canadians compared to North Shore.

We also described inbreeding among LCAs. Only 13% of pairs of related individuals had one inbred LCA or more but this percentage varied greatly from one population to another. The proportions of pairs with at least one inbred LCA was more than half for the Saguenay, North

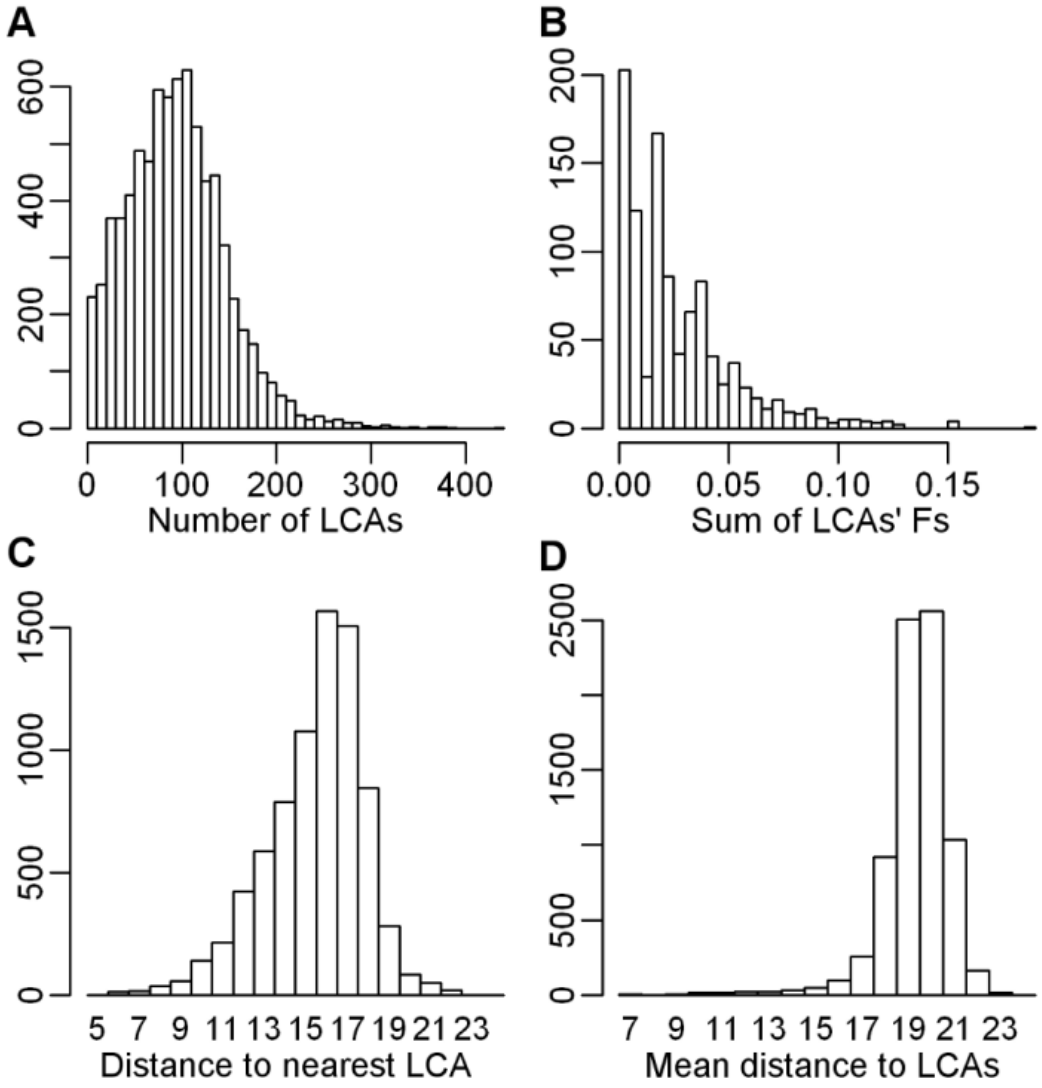

Figure 2.1 Distributions of genealogical characteristics

Histograms of genealogical characteristics calculated for each of the 7704 related pairs (i.e. genealogical kinship > 0). A) Number of lowest common ancestors (LCAs); B) Sum of LCAs' inbreeding coefficients (Fs) among pairs having at least one inbred LCA (n=1034); C) Distance to nearest LCA; D) Mean distance to LCAs.

Shore and Acadian populations, 31% for Gaspesian French Canadians and less than 8% for the other populations. The number of inbred LCAs for a pair of individuals with LCAs ranged from 0 to 13 and inbreeding coefficients ranged from 0.00006 to 0.06, which are approximately equivalent to individuals with parents that are seventh degree relatives and first cousins, respectively. Figure 2.1B (p.57) shows the sum of all LCAs' inbreeding coefficients, which is the measure that we chose to summarize the inbreeding information. This sum ranged from 0.0001 to 0.2 for pairs of individuals with at least one inbred LCA. Overall, distributions of genealogical characteristics reflect the diversity and complexity of the relationships present in the structured founder population of Quebec.

Comparison of different IBD sharing detection methods

Before comparing results from selected methods, we looked at results from the only method using phased data, GERMLINE, for which we used two different phasing methods (ShapeIT and Beagle). Haplotypes obtained with different phasing methods are not consistent and this might impact IBD inference. Indeed, data phased with ShapeIT provided IBD results that were more strongly correlated with the genealogical information than those phased with Beagle. The correlation between total length of IBD segments and genealogical kinship coefficients for results from GERMLINE was 0.92 for genotype phased with ShapeIT and 0.72 for genotype phased with Beagle. Hence, we retained GERMLINE's results with ShapeIT phasing for further analyses.

In the whole sample from the Province of Quebec, we observed Pearson's correlation coefficients ranging from 0.69 to 0.92 for the total length of IBD sharing identified with the different methods against the genealogical kinship coefficient (Table 2.1 p. 59, Supplementary Figure 2.6 p.70). Three methods (GERMLINE, FastIBD and IBDLD) stand out with correlation coefficients of 0.92. IBD sharing identified by PLINK and SLRP was less concordant with genealogical information.

Table 2.1 Pearson's correlation coefficients between total length of IBD sharing and kinship coefficients for each population and each method

Population	*Methods*				
	PLINK	*GERMLINE*	*FastIBD*	*IBDLD*	*SLRP*
ACA	0.87	0.88	0.89	0.89	0.85
GFC	0.92	0.91	0.92	0.92	0.88
LOY	-0.03	0.84	0.86	0.86	0.01
MON	0.10	0.39	0.46	0.45	-0.02
NS	0.85	0.88	0.90	0.88	0.83
QUE	0.09	0.31	0.45	0.42	0.15
SAG	0.15	0.82	0.84	0.83	0.13
PQ	0.77	0.92	0.92	0.92	0.69

Abbreviations: ACA, Acadians; GFC, Gaspesian French Canadians; LOY, Loyalists; NS, North Shore; MON, Montreal; QUE, Quebec City area; SAG, Saguenay; PQ, Whole sample from the Province of Quebec.

To get a better idea of which method provided the most appropriate results for our data, we further examined the correlation between IBD sharing and kinship in each sub-population separately. Correlations varied across populations (Table 2.1). We noted the low correlation of IBD sharing inferred by some methods in the Saguenay region with kinship coefficient despite the presence of a noteworthy degree of relatedness among individuals in this region. Less surprisingly, populations with lower expected relatedness, such as Montreal and Quebec City areas, had lower correlations with values ranging from 0.02 to 0.46. We also noted that correlations found for the Loyalists were either very good (0.84 to 0.86) or very weak (-0.03 or 0.01).

Results from FastIBD were retained for further analyses. Assuming that genealogical kinship is the true expected kinship, FastIBD was among the fastest (see Supplementary Table 2.4 p.74 for detailed information on computation time) and best reflected the relatedness described by our genealogical data, as evaluated by the correlation between total length of IBD sharing and genealogical kinship coefficient.

Genealogical measures versus inferred IBD sharing

Before looking at the relationship between IBD sharing and different genealogical variables, we examined the impact of genealogical completeness on the correlation between total length of IBD sharing and the genealogical kinship coefficient. We recalculated the correlation coefficients with pairs of individuals having more than 50% of their genealogical information complete at the 5^{th} generation and also with the same completeness at the 10^{th} generation. Almost no change was observed at the 5^{th} generation, while at the 10^{th} changes in correlation coefficients were small (0 to 0.10, all within one standard deviation of the estimates) except for the Loyalists that did not have enough complete pairs at the 10^{th} generation to recalculate the correlation. We chose to keep all pairs in our sample.

As IBD sharing was highly correlated with genealogical kinship coefficient, a simple linear regression fits the data well. Hence, the overall degree of relatedness is well captured by overall IBD sharing with 85% of the variance in kinship coefficients explained by total length

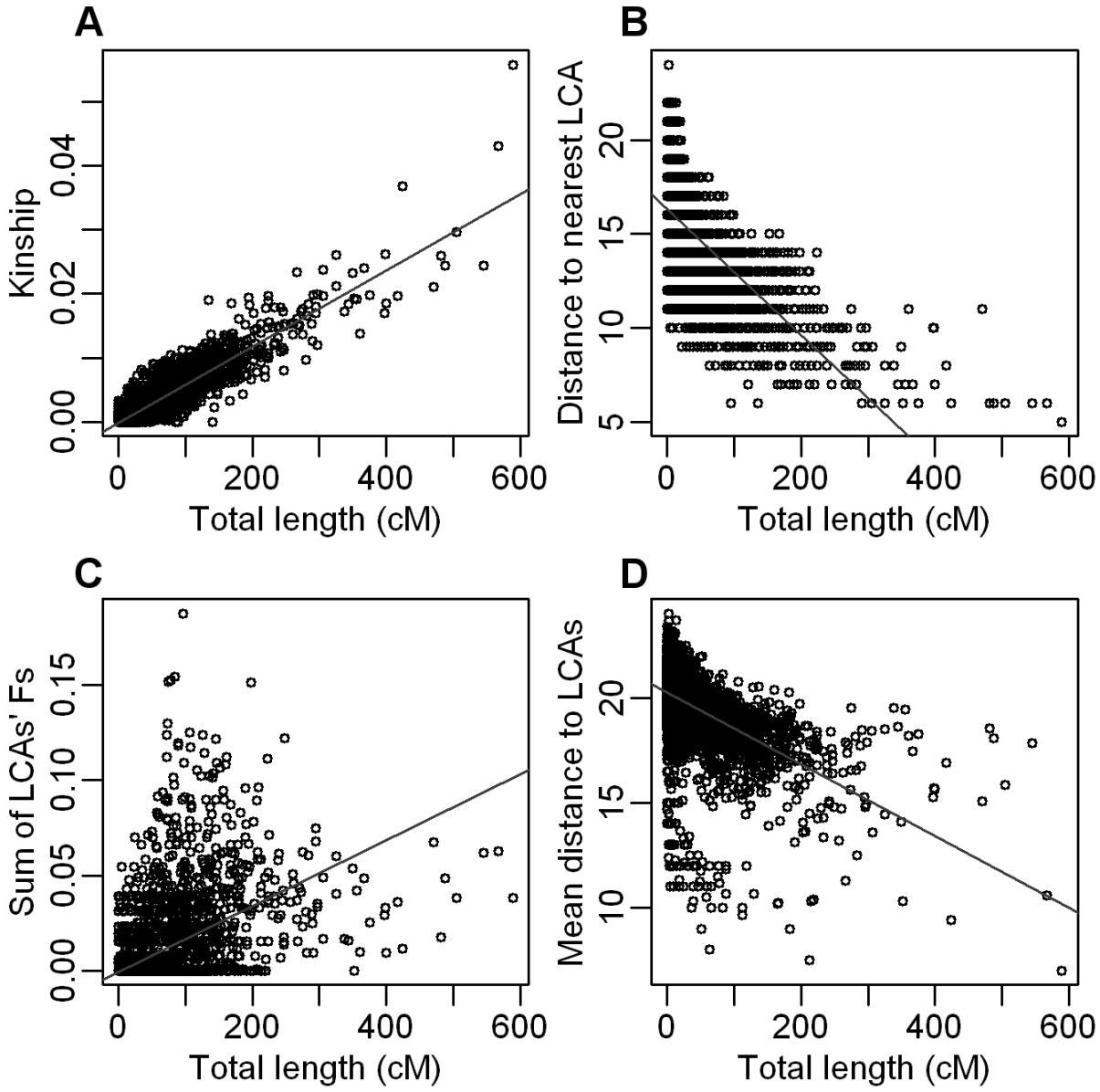

Figure 2.2 IBD sharing and genealogical characteristics

Scatter plots of total length of IBD sharing versus genealogical characteristics for each pair (n=9730): A) kinship coefficients; B) distance to nearest lowest common ancestor (LCA); C) sum of LCAs' inbreeding coefficients (Fs); and D) mean distance to LCAs. A simple linear regression line is plotted in grey on each graph.

of IBD sharing (Figure 2.2A p.61). Total length of IBD sharing also reflected characteristics of relatedness, such as shorter distance to LCA or having an inbred ancestor. Total length of IBD sharing explained 26% of the variance in the mean distance to LCAs, 39% of the variance in the distance to the nearest LCA, and 31% of the variance in the sum of LCAs' inbreeding coefficients (Figure 2.2B-D p.61). As pairs of individuals sharing an inbred common ancestor seemed to be a distinct group we separated the whole sample based on this criterion to assess the impact on IBD sharing. Comparing the two groups obtained, we observed significantly more IBD sharing for pairs having at least one inbred LCA (Supplementary Figure 2.7 p.71). These pairs have, on average, 7.2 times more total length of IBD sharing and 4.5 times more IBD segments.

IBD sharing in populations

The amount of IBD sharing per population is shown on Figure 2.3 (p.63). Each dot represents the mean number of segments shared per pair for specific length ranging from 2 to 15 cM. The number of segments and their length vary with the degree of relationships, yielding distinct curves for the different levels of kinship present in the populations. The Acadians, which have the highest levels of kinship, have a curve well above the other populations. The Saguenay and North Shore curves overlap, reflecting similar kinship levels in these two populations. Montreal and Quebec City show lower and more variable levels of IBD sharing. We also observed a clear difference between our whole sample from Quebec and the HapMap CEU sample. On average pairs of individuals from Quebec shared 3.8 IBD segments and have 21.3 cM of IBD sharing while those from HapMap CEU share 2.7 IBD segments and have 8.0 cM of IBD sharing. Thus, pairs of individuals from Quebec also shared longer segments, with segments smaller than 5 cM representing 63% and 96% of segments for Quebec and HapMap CEU, respectively.

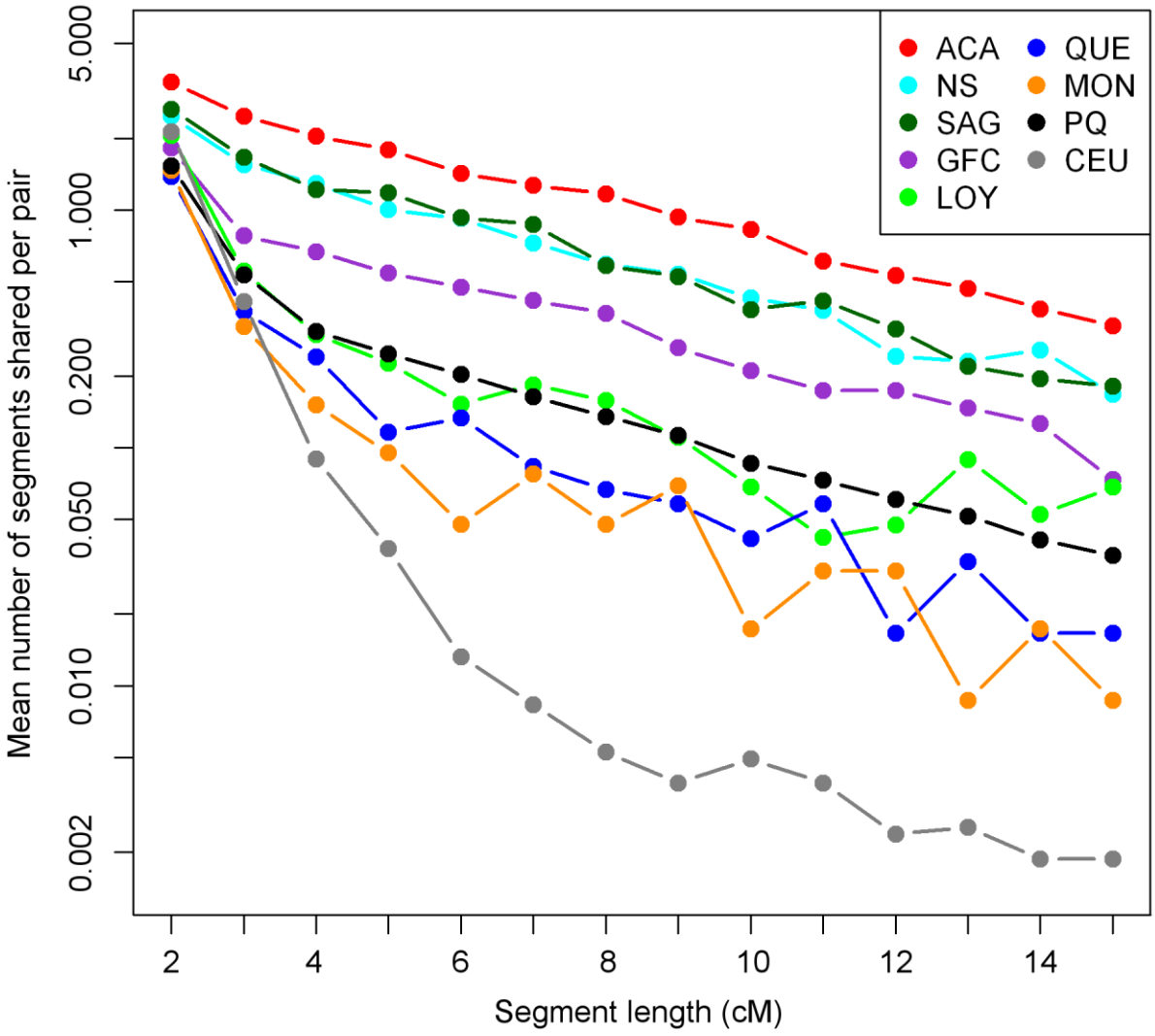

Figure 2.3 Pairwise IBD sharing in each population

The mean number of segments shared is shown (y-axis, log-scale) per pair for specific 1 cM class length ranging from 2 to 15 cM. ACA, Acadians; GFC, Gaspesian French Canadians; LOY, Loyalists; NS, North Shore; MON, Montreal; QUE, Quebec City area; SAG, Saguenay; PQ, Whole sample from the Province of Quebec; CEU HapMap.

Whole-genome IBD sharing

Figure 2.4 (p.65) shows the proportions of pairs of individuals having IBD sharing at specific chromosomal positions across the whole genome. Patterns across populations are different and, as in Figure 2.3, we can see that average sharing differs among populations, with the CEU sharing less than the Quebec population. Some IBD sharing seems consistent across populations for example around the HLA region on chromosome 6 where a peak can be observed for the whole Quebec sample and CEU sample.

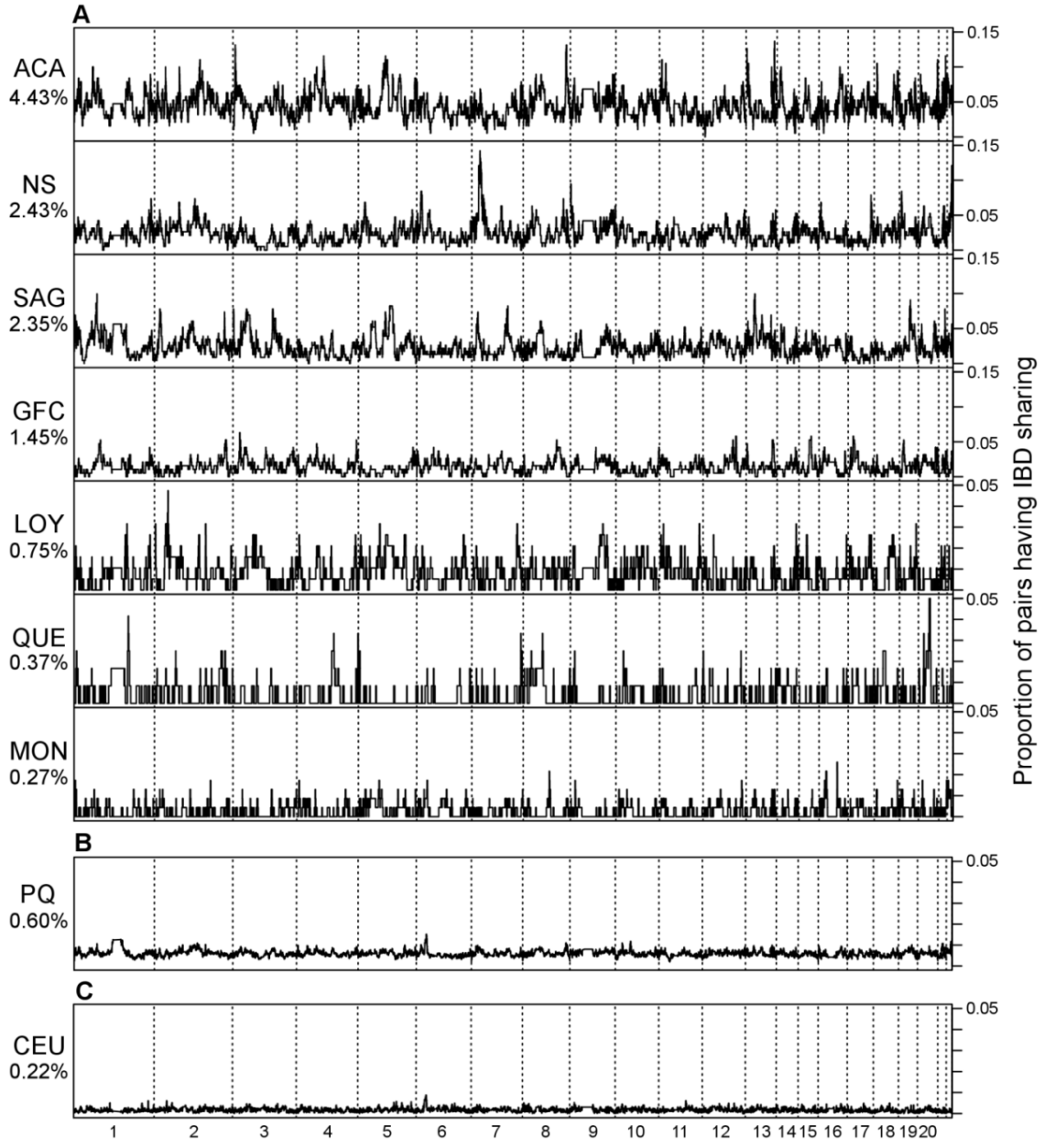

Figure 2.4 Genome-wide patterns of IBD sharing in each population

The proportion of IBD sharing across the genome is shown for each population with vertical dashed lines to separate chromosomes and mean proportion of sharing indicated on the left. The proportion was calculated at each position that was genotyped in our data. Note that the scale is different for the first four graphs in panel A. A) Quebec populations: ACA, Acadians; GFC, Gaspesian French Canadians; LOY, Loyalists; NS, North Shore; MON, Montreal; QUE, Quebec City area; SAG, Saguenay; B) PQ, Whole sample from the Province of Quebec; C) CEU HapMap.

Discussion

In this study, we first compared IBD inference provided by five different methods by correlating total length of IBD sharing with genealogical kinship. To our knowledge, our study is the first to provide a comparison of the performance of different methods in a complex dataset from a founder population. It is difficult to evaluate the performance of methods in a real-data setting since we do not know the truth. Because of the availability of extended genealogies in our population, we could evaluate, at least in part, the performance of IBD detection methods by comparing them to genealogical information. Our results confirmed the importance of a well-defined and flexible model or algorithm for IBD inference and identified FastIBD, GERMLINE and IBDLD as the best-performing methods based on the high correlations between total length of IBD segments shared by pairs of individuals and their genealogical kinship coefficient. As noted in previous studies, simple models that do not consider genotyping errors and LD, such as that implemented in PLINK, yield lower resolution of IBD detection (Gusev *et al.* 2008; Browning and Browning 2010). In our sample, the smallest segment detected with PLINK was 3.8 cM long, almost twice as our threshold of 2 cM. With respect to genotyping errors, a modification of PLINK has been proposed in order to include genotyping confidence scores into the IBD inference process which could improve IBD inference (Markus *et al.* 2011). SLRP has previously been shown to yield a high accuracy of IBD detection compared to GERMLINE and FastIBD in simulated data (Palin *et al.* 2011). However, in our population, SLRP identified more IBD sharing than the other methods while yielding the lowest correlation coefficients with genealogical kinship overall.

We selected FastIBD for further analyses because IBD sharing within populations was more associated with genealogical information with this method and it was fast to run. We recognize that we did not optimize the parameters selected for each method but simply used the ones recommended by the authors for their methods. Parameter optimization could have affected our comparison and improved our results. However, our results are consistent with most simulations reported in the literature comparing different methods. We also restricted our study to five methods since other existing methods were more difficult to use or not

implemented in a software (Albrechtsen *et al.* 2009; Moltke *et al.* 2011; Stevens *et al.* 2011; Brown *et al.* 2012).

Using results from FastIBD, we then related IBD sharing to the different genealogical measures. Total length of IBD sharing explained a large portion of the variance in kinship coefficients. Our results highlight the variability in realized IBD sharing for a variety of pairs of remotely related individuals with known kinship. Not surprisingly, total length of IBD sharing also explained more of the variance in the distance to nearest LCAs than of the variance in mean distance to LCA since the most recent ancestors have a higher impact on IBD sharing. We aggregated inbreeding coefficients from LCAs into a unique sum and found that IBD sharing also explained a noteworthy part of its variance. However we are conscious that a pair of individuals could have no inbred LCA identified but still share a more distant inbred ancestor. This occurred for only 20 pairs of individuals. Some shared inbred LCAs may also not be identified because of lower genealogical completeness. This might explain a few pairs of individuals without inbred LCAs (shown as outliers on Supplementary Figure 2.7 p.71) that had an important amount of IBD sharing compared to their group average and that had inbred ancestors that were not shared according to the information available.

Length of segments identified in the different populations was also a good way to identify population differences. The odds of sharing more segments as well as longer segments were higher in population with more relatedness, as expected. Furthermore, IBD sharing in the whole sample was very high and, as expected, mean length of segments inferred (data not shown) was higher than in any other HapMap or Ashkenazi Jewish populations considered in Gusev *et al*. except for one sample in which many pairs were closely related (closer than 1st degree cousin according to IBD inference) (Gusev *et al.* 2012). The high IBD sharing and increased proportion of longer segments is explained by founder events that occurred and population expansions following them (Palamara *et al.* 2012). The fact that the Saguenay population size underwent a 25-fold increase in only a century, from 1861 to 1961, while the whole Quebec population increased about 5 times, is in good conjunction with our results.

Despite important differences in mean proportion of IBD sharing between Quebec and HapMap CEU, we noted the presence of a common peak of IBD on chromosome 6 covering the HLA region. Increased IBD sharing has been reported for this region in several dpopulations and could be the results of selection (Albrechtsen *et al.* 2010a; Gusev *et al.* 2012). As pointed by Browning and Browning (Browning and Browning 2012), IBD inference will be facilitated in region with high LD but LD may also lead to overestimating the true IBD sharing relative to recent common ancestor. Knowing that important LD normally arises in presence of natural selection, the relevance of an excess of IBD sharing in the HLA region should be investigated more deeply.

IBD detection methods are useful in many contexts such as identifying phasing errors or polymorphic deletions, estimating heritability (Gusev *et al.* 2008; Price *et al.* 2011), inferring kinship (Huff *et al.* 2011; Stevens *et al.* 2011, 2012; Henn *et al.* 2012; Thornton *et al.* 2012) and mapping diseases in association studies (Gusev *et al.* 2011; Browning and Thompson 2012). In cases where genealogical information is not available we now know that IBD is an alternative to account for unknown relatedness (Newman *et al.* 2001). Even in samples that are widely used such as those coming from CEPH, precautions are necessary as important consanguinity as been identified recently (Stevens *et al.* 2012). Our study is an additional example putting forward the importance of considering relatedness in a sample before studying it. The high correlation that we observed between genealogical information and IBD sharing, over the wide range of remote relatedness present in our study population, further demonstrates the usefulness of genomic IBD detection to capture even complex relatedness involving inbreeding and our findings can guide the interpretation of results in other population without genealogical data. Our study highlights the great variety in types of relatedness present in the French Canadian founder population and how this complex relatedness is reflected in patterns of IBD sharing. Using these patterns, it is thus possible to gain insight on the types of distant relatedness in a sample from a founder population like Quebec, leading to better genetic study design and analysis.

Supplementary Figures and Tables

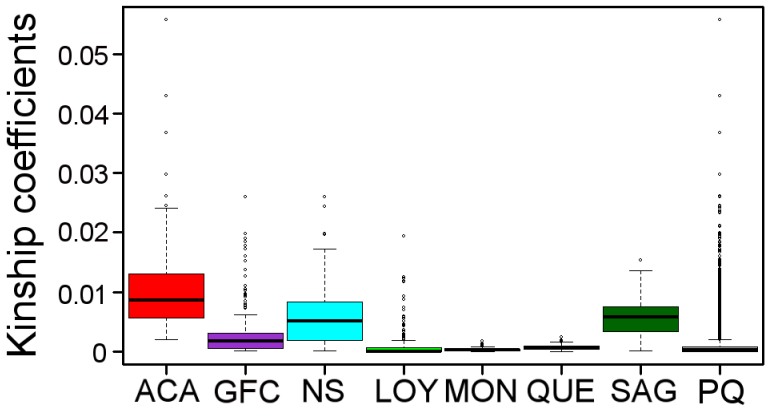

Figure 2.5 Boxplot of genealogical kinship coefficients for each population

ACA, Acadians; GFC, Gaspesian French Canadians; LOY, Loyalists; NS, North Shore; MON, Montreal; QUE, Quebec City area; SAG, Saguenay; PQ, Whole sample from the Province of Quebec.

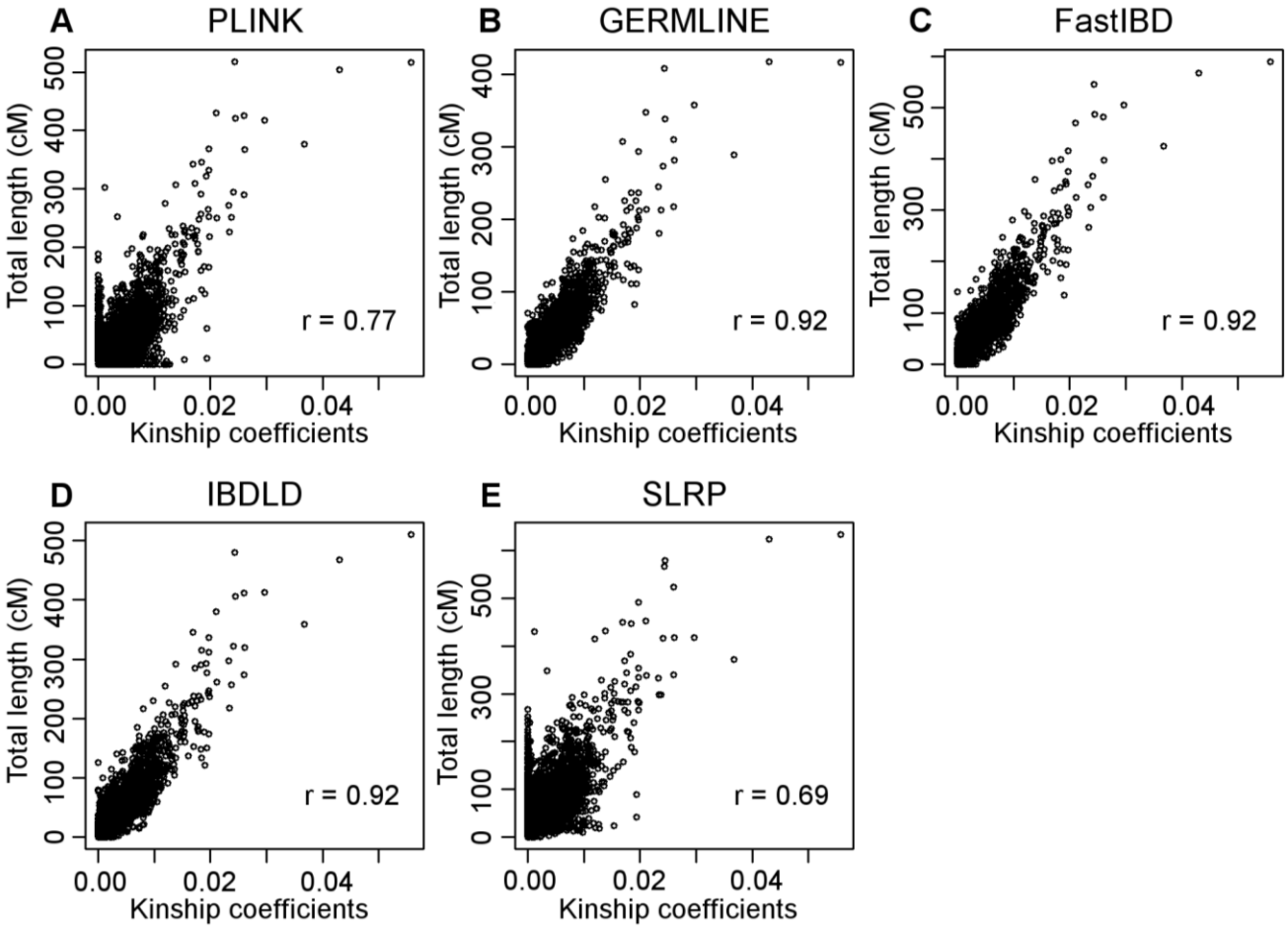

Figure 2.6 Scatterplots of total length of IBD sharing versus kinship coefficients for the whole sample

Each plot shows on the y-axis the total length of IBD shared by pairs of individuals and inferred by a method : A) PLINK; B) GERMLINE; C) FastIBD; D) IBDLD; E) SLRP. The total length of IBD sharing is related to the kinship coefficient, on the x-axis calculated from the genealogical information. The Pearson's correlation coefficient between variables is presented in the bottom-right corner of each plot.

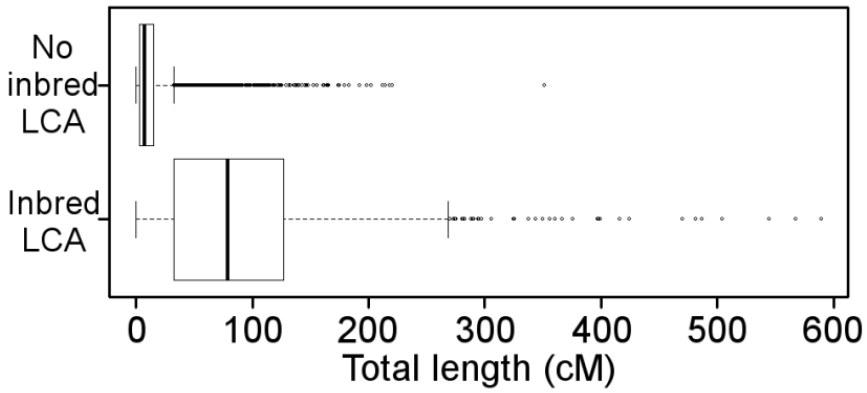

Figure 2.7 IBD sharing and inbreeding

Boxplots of total length of IBD sharing in pairs of individuals having at least one inbred LCA or none.

Table 2.2 List of the 109 HapMap CEU samples used

NA06984	NA11839	NA12043	NA12287	NA12763
NA06985	NA11840	NA12044	NA12340	NA12775
NA06986	NA11843	NA12045	NA12341	NA12776
NA06989	NA11881	NA12056	NA12342	NA12777
NA06993	NA11882	NA12057	NA12343	NA12778
NA06994	NA11891	NA12144	NA12347	NA12812
NA07000	NA11892	NA12145	NA12348	NA12814
NA07022	NA11893	NA12146	NA12383	NA12815
NA07031	NA11894	NA12154	NA12399	NA12827
NA07037	NA11917	NA12155	NA12400	NA12828
NA07051	NA11918	NA12156	NA12413	NA12829
NA07055	NA11919	NA12234	NA12489	NA12830
NA07056	NA11920	NA12239	NA12546	NA12842
NA07345	NA11930	NA12248	NA12716	NA12872
NA07346	NA11931	NA12249	NA12718	NA12873
NA07347	NA11992	NA12264	NA12748	NA12874
NA07357	NA11993	NA12272	NA12749	NA12875
NA07435	NA11994	NA12273	NA12750	NA12889
NA11829	NA11995	NA12275	NA12751	NA12890
NA11830	NA12003	NA12282	NA12760	NA12891
NA11831	NA12005	NA12283	NA12761	NA12892
NA11832	NA12006	NA12286	NA12762	

Table 2.3 IBD inference methods used and options (default or not) specified for each one

Method and version	Data	Genotyping error rate	Other parameters specified or data manipulations made
PLINK 1.07	Unphased dataset pruned to remove LD	Not allowed	- All pairs included (no inclusion/ exclusion threshold on genome-wide IBD)
GERMLINE 1.5.0	Phased by either Beagle or Shape-It	Homozygous missmatch 2 (heterozygous 0)	- Extension on haplotypes - Minimum length for match 1 cM
FastIBD 3.3.1	Unphased	-	- Performed a fastIBD analysis 10 times and combined results - IBD criteria: fastIBD score $< 10^{-10}$ and segment length is ≥ 1 cM
IBDLD 2.06	Unphased	0.01	- Method used : GIBDLD - Extension of the HMM to include LD by conditioning on the 10 previous markers with a ridge regression.
SLRP 1.0-48-g20ad5c4	Unphased	0.001	

Table 2.4 Runtime comparison for the different IBD inference methods

For chromosome 12[1] (26 668 SNPs). All experiments were conducted on a Linux node of 1 X 2.66 GHz Xeon CPU with 2 GB of memory.

Methods	Data	Time (hr:min:sec)
PLINK	Unphased	00:06:41
GERMLINE	Phased	00:00:01[2]
FastIBD	Unphased	00:42:11[3]
IBDLD	Unphased	11:35:41
SLRP	Unphased	00:46:57

[1] The chromosome 12 is selected as an example, since it has a number of SNPs halfway in between the smallest and the longest chromosome.

[2] When adding the phasing step time we obtained 2:22:11 and 00:04:29 when phasing with Shape-It and Beagle, respectively.

[3] FastIBD recommends combining results from 10 runs/chromosome. If these processes are parallelized, time decreases to 00:04:20.

Chapter 3:
GENLIB: an R package for the analysis of genealogical data

Héloïse Gauvin, Jean-François Lefebvre, Claudia Moreau, Eve-Marie Lavoie, Damian Labuda, Hélène Vézina, and Marie-Hélène Roy-Gagnon

Reference: Gauvin H, Lefebvre J-F, Moreau C, Lavoie E-M, Labuda D, Vézina H, Roy-Gagnon, M-H. GENLIB: an R package for the analysis of genealogical data. *BMC Bioinformatics*, 2015; 16; 160. doi: 10.1186/s12859-015-0581-5

Authors' contribution

In this paper, my contribution is:

- The analysis of the genealogical corpus;
- The simulation study and its analysis;
- The implementation of new functions with JFL;
- The package testing with CM, JFL and EML;
- The design of the study with MHRG, DL and HV;
- Writing of the paper.

Contributions of other authors are: JFL built the R package and provided bioinformatic support. HV provided the genealogical datasets. MHRG oversaw the making of the R package and the data analysis. All authors reviewed and approved the manuscript.

Acknowledgements

I am grateful to Louis Houde who designed the first version of GENLIB using S-Plus with programming contributions from Jean-Luc Allard, Claude Bhérer and Valéry Roy-Lagacé. I also thank all the anonymous reviewers and editors for their valuable comments on the analyses, which helped improve this manuscript.

This work was supported by a grant from the Fonds de la Recherche en Santé du Québec (FRSQ) to MHRG. Support from the FRSQ Réseau de Médecine Génétique Appliquée (RMGA) and the Canadian Institutes of Health Research (CIHR; to DL and HV) is also gratefully acknowledged. HG is the recipient of a scholarship from the RMGA.

Abstract

Background: Founder populations have an important role in the study of genetic diseases. Access to detailed genealogical records is often one of their advantages. These genealogical data provide unique information for researchers in evolutionary and population genetics, demography and genetic epidemiology. However, analyzing large genealogical datasets requires specialized methods and software. The GENLIB software was developed to study the large genealogies of the French Canadian population of Quebec, Canada. These genealogies are accessible through the BALSAC database, which contains over 3 million records covering the whole province of Quebec over four centuries. Using this resource, extended pedigrees of up to 17 generations can be constructed from a sample of present-day individuals.

Results: We have extended and implemented GENLIB as a package in the R environment for statistical computing and graphics, thus allowing optimal flexibility for users. The GENLIB package includes basic functions to manage genealogical data allowing, for example, extraction of a part of a genealogy or selection of specific individuals. There are also many functions providing information to describe the size and complexity of genealogies as well as functions to compute standard measures such as kinship, inbreeding and genetic contribution. GENLIB also includes functions for gene-dropping simulations.

The goal of this paper is to present the full functionalities of GENLIB. We used a sample of 140 individuals from the province of Quebec (Canada) to demonstrate GENLIB's functions. Ascending genealogies for these individuals were reconstructed using BALSAC, yielding a large pedigree of 41,523 individuals. Using GENLIB's functions, we provide a detailed description of these genealogical data in terms of completeness, genetic contribution of founders, relatedness, inbreeding and the overall complexity of the genealogical tree. We also present gene-dropping simulations based on the whole genealogy to investigate identical-by-descent sharing of alleles and chromosomal segments of different lengths and estimate probabilities of identical-by-descent sharing.

Conclusions: The R package GENLIB provides a user friendly and flexible environment to analyze extensive genealogical data, allowing an efficient and easy integration of different types of data, analytical methods and additional developments and making this tool ideal for genealogical analysis.

Background

Studying founder or isolated populations is a major asset in reducing or taking into account the heterogeneity of genetic and environmental factors involved in complex diseases. Indeed, potential advantages of founder populations include greater genetic and environmental homogeneity and in some instances, the availability of extensive genealogical records (Bourgain and Génin 2005). Many founder populations across the world are currently studied, among others Iceland (Gulcher *et al.* 2001), the Amish (McKusick 1978; Agarwala *et al.* 2003; Morton *et al.* 2003; Roy-Gagnon *et al.* 2008), the Hutterites (Newman *et al.* 2001), the Mormons (Slattery and Kerber 1993), Finland (Peltonen *et al.* 1999; Uimari *et al.* 2005), Newfoundland (Rahman *et al.* 2003), the Sardinians (Ciullo *et al.* 2006) and the French Canadian founder population of Quebec (Vézina *et al.* 2005a; Moreau *et al.* 2011, 2013). As the value of genealogical resources is increasingly recognized (Kristiansson *et al.* 2008; Larmuseau *et al.* 2013a), efforts have been made to obtain extended genealogical information and build genealogical databases for use in genetic studies. To our knowledge, there are two available, freestanding packages for genealogical analysis, named PEDSYS (Dyke 1999) and PedHunter (Agarwala *et al.* 1998; Lee *et al.* 2010). However, each lacks some functionality that we have found useful in our studies of the Quebec population and neither PEDSYS nor PedHunter can be used within a statistical computing environment such as R.

The BALSAC database is an example of a large genealogical database that has proven very valuable for genetic and demographic research in the Quebec founder population (Scriver 2001; Laberge *et al.* 2005b; QRS 2015). Briefly, the arrival of French settlers in the province of Quebec, Canada, at the beginning of the 17th century followed by the British Conquest about a century and a half later shaped the colonization history of Quebec and led to successive regional founder effects (Gagnon and Heyer 2001; Scriver 2001). From ~8,500 settlers, the French speaking population now represents about 80% of the almost 8 million inhabitants of Quebec (Statistics Canada 2012). This fast population expansion and the regional founder effects have shaped genetic variation and population structure in Quebec. The related demographic information can be accessed using the BALSAC database, created 40

years ago, which now includes over 3 million vital event records providing information on 5 million individuals (BALSAC 2014b). These records have been linked allowing for the reconstruction of ascending genealogies from present-day individuals going back over four centuries. The BALSAC data access policy and the procedures to request access to the BALSAC genealogical data can be found on the BALSAC website (BALSAC 2014b).

Specialized software is required in order to optimally exploit the full potential of these well preserved, exhaustive and detailed genealogical data. The S-Plus library of functions GENLIB was originally developed to work with the BALSAC database to manage, describe and visualize genealogies as well as perform allele-dropping simulations. The GENLIB S-Plus library has been used to study the selective advantage to be on an expanding wave front during a range expansion (Moreau *et al.* 2011), the genetic contribution of non-French groups to the current population (Tremblay *et al.* 2008; Moreau *et al.* 2013), population structure (Bherer *et al.* 2011; Roy-Gagnon *et al.* 2011), drug intolerance for an inherited disorder in multigenerational pedigree (Tremblay *et al.* 2014) and patterns of genetic sharing relatively to expected sharing from genealogical information (Gauvin *et al.* 2014).

We have extended and implemented GENLIB in R to provide a freely accessible version of the software within a user-friendly and flexible environment and to allow a more efficient and easier integration of different types of data and analytical methods, also facilitating future developments. Such flexibility is especially important in the context of integrating large-scale genomic data and genealogical data. While other pedigree software exist, few are comprehensive, free and specific to extended genealogical databases. For example, the PedHunter (Agarwala *et al.* 1998; Lee *et al.* 2010) and PEDSYS (Dyke 1999) software include many functions to describe and analyze genealogical data but do not provide gene-dropping simulation functions. In this article, we present most of the functions in the GENLIB package. Using these functions, we perform a detailed description of a large genealogical corpus including 41,523 individuals provided as an example dataset in the R package. Then we show an example of gene-dropping simulations, a procedure in which hypothetical alleles, or chromosomal segments, are assigned to ancestors and dropped down the whole genealogical tree (MacCluer *et al.* 1986).

Implementation

Overview

GENLIB was originally written in the S programming language, integrating C++ functions to accelerate and extend S programming. We started with the same C++ functions in the translation of GENLIB into an R package and we extended both the C++ and R translation of the S code. Interaction between R and C++ is well supported and high quality R libraries are available to help implement such packages. The package is portable across Linux-like and Windows operating systems, and benefits from the advantages of the R environment, such as being open-source, available online under the GNU General Public License, and offering a wide range of complementary functions. All code has been tested using version 3.1.0 of R (http://cran.r-project.org). Version 1.0 of the GENLIB package presented in this paper includes over 40 functions. GENLIB is included in the CRAN packages repository (http://cran.r-project.org/web/packages) and is also available on the BALSAC website (http://balsac.uqac.ca). By installing the package, users also have access to the genealogical data described in this paper.

Functions implemented

The starting point for any genealogical analysis in GENLIB is the creation of a genealogical object using the *gen.genealogy* function. The following input information needs to be provided in a matrix or data frame where each line corresponds to a subject: the subject identification number (ID), the subject's father ID, the subject's mother ID and subject's sex (coded 1/2 for male/female). All values must be numerical. Functions from the package are grouped into 4 categories: i) management, ii) description and visualisation, iii) computation and iv) simulations. Table 3.1 provides an overview of most functions included in GENLIB.

The first category includes functions allowing to track specific individuals (founders, siblings, probands, etc) or to extract a part of a genealogy or a lineage. New functions to identify

Table 3.1 Overview of GENLIB functions

Name	Use
Functions to manage genealogical objects	
gen.genalogy	To create a basic genealogical object
gen.lineages	To extract parental lineages from a genealogical object
gen.branching	To extract a subset of a genealogical object
gen.genout	To output a genealogical object as a data frame
gen.founder, gen.half.founder, gen.pro, gen.parent, gen.sibship, gen.children, gen.findFounders, gen.findMRCA	To identify specific individuals (founder, half-founder, proband, parent, sibship, children, common founder, most recent common ancestor)
Functions to describe…	
gen.nomen, gen.nowomen, gen.noind, gen.nochildren	...the number of men, women or individuals in a genealogy and number of kids an individual has
gen.min, gen.mean, gen.max	...the minimal, mean or maximal generation at which an individual can be found in a genealogy (first generation is coded 0)
gen.depth	...the number of generations in the genealogy
gen.completeness	...genealogical data completeness
gen.rec	...how many individuals within the specified individual group descend from each specified ancestor
gen.occ	...how many different (but not mutually exclusive) paths link an ancestor to a descendant
gen.meangendepth	...how much rooted are the genealogical lineages
gen.implex	...the extent of pedigree collapse within an individual's genealogy
gen.findDistance	…distance between individuals through a specific ancestor
gen.find.Min.Distance.MRCA	…the shortest distances between individuals
Functions to plot…	
gen.graph	...the genealogy
Functions to compute…	
gen.phi	...the kinship matrix at specified generations
gen.f	...the inbreeding coefficients at specified generations
gen.gc	...the genetic contribution of ancestors to individuals
Gene-droppping simulation functions	
gen.simuProb	To compute the probability that individuals have 0, 1 or 2 copies of a disease allele knowing how many their ancestors had

gen.simuSample, gen.simuSampleFreq	To obtain the number (frequencies) of disease alleles for each individual taking into account each ancestor's carrier status
gen.simuSet	As function gen.simuSample with option to customize transmission probabilities according to the parent's and/or subject's sex

Note: Additional functions (e.g., to calculate the variance associated with kinship and other measures) are available but not included in the table

common founders and most recent common ancestors (MRCA) to a group of individuals have been implemented. In the second category, there are functions to count particular individuals, to compute the generational completeness and to describe the depth of the genealogical information and the number of paths linking two individuals. The completeness at a given generation (g) is the proportion of ancestors present in the genealogy (A_g) relative to the maximum possible number of ancestors, i.e. assuming that all individuals at the previous generation have two known parents (2^g) (see Table 3.2, p.85, for all formulas). When reproduction between two related individuals happens, the number of distinct ancestors in the family tree of their offspring is reduced. The implex index, which can also be computed with GENLIB, quantifies this collapsing phenomenon. In other words, the implex measures how much an individual's ancestor tree deviates from a binary tree where the individual, his/her 2 parents, 4 grandparents, 8 great-grandparents and so on are all distinct individuals. Formally, the implex is calculated by taking the actual number of distinct ancestors (i.e., each ancestor is counted only once regardless of its number of occurrences) relative to the theoretical number of ancestors (2^g) at a specified generation *g* (Cazes and Cazes 1996). Therefore the implex is also influenced by the information known about the ancestors and is always bounded by the completeness as an upper limit. In fact, to be informative, the implex should always be interpreted in the context of the completeness. Finally, the mean genealogical depth can be computed to get the expected generation where the founders can be found (Kouladjian 1986; Cazes and Cazes 1996). This value reflects the amount of information available in the genealogy as it is the average length of each lineage. In addition to characterizing genealogies, GENLIB provides a function to plot the genealogy (see Supplementary Figure 3.5 and Figure 3.6 for examples of graphs, pp. 100-101).

Table 3.2 Formulas of genealogical measures in GENLIB

Completeness	$C_g = \frac{A_g}{2^g}$	A_g :number of known ancestors at generation g F_g : number of founders at generation g T_g : number of expected ancestors at generation g ($=2^g$) N_f : generation of founder f (summation over all generations g or over all founders f)
Implex index	$I_g = \frac{Distinct\ A_g}{2^g}$	
Mean genealogical depth	$D = \sum_g g\frac{F_g}{T_g} = \sum_f N_f \frac{1}{2^{N_f}}$	
Variance of mean genealogical depth	$\sum_g g^2 \frac{F_g}{T_g} - D^2$	
Kinship	$\varphi_{ij} = \sum_a (1+f_a)\frac{1}{2}^{g_{ai}+g_{aj}+1}$	g_{ai} (g_{aj}) : number of generations between subject i (j) and ancestor a, common to i and j (summation over all lines of descent of ancestor a)
Inbreeding	$f_i = \varphi_{kl}$	k, l : the parents of i
Genetic contribution	$GC(s,a) = \sum_p \left(\frac{1}{2}\right)^{g_p}$	g_p : number of generations between subject s and ancestor a through path p (summation over all possible paths p)

The third category of functions includes demogenetic functions. It is possible to compute the kinship and inbreeding coefficients (Malécot 1948; Karigl 1981; Thompson 1986) and the genetic contribution of an ancestor to its descendants (Roberts 1968; O'Brien *et al.* 1988) (see Table 3.2 for formulas). For kinship calculations that can be computationally intensive with GENLIB, the kinship function allows multi-threading. We compared kinship computation times between GENLIB and PedHunter by calculating kinship coefficients for all pairs of 140 probands using a six-core processor running at 2.667 GHz with 12 GB of RAM. The same results were obtained in 2 minutes with the multi-threading option in GENLIB and in 1 minute with PedHunter (excluding the preliminary data re-formatting steps required by PedHunter). The genetic contribution is the calculation of an ancestor's contribution to the genetic makeup of an individual and this is obtained by summing up the probabilities of transmission over all genealogical paths connecting an ancestor and its descendant. Like other functions in the package, the computation of genetic contribution is customizable, meaning that it is possible to modify the inheritance pattern, which assumes that regardless the sex of parents and child, half the genetic material is transmitted.

Lastly, GENLIB provides functions to perform gene-dropping simulations, simulations which have already been proven to be very valuable to study BALSAC genealogies (Heyer 1999; Tremblay *et al.* 2003). Other software implementing gene-dropping simulations (such as Mendel (Lange *et al.* 2013)), were conceived with a focus on whole pedigrees as opposed to reconstructed ascending genealogies from large founder populations, resulting in different functionalities and usage compared to GENLIB such as, for example, the ability to easily restrict simulations to specific individuals or ancestors. In GENLIB, different functions and their options allow to simulate genotype data for specific individuals in the genealogy based on a provided genealogical structure and according to the number of alleles carried by specific ancestors. Briefly, genotypes can be assigned to selected ancestors and segregated down the genealogy paths (MacCluer *et al.* 1986). We added functionality to the gene-dropping simulations by implementing a segment-dropping option where the user is able to specify a recombination probability, i.e., a chance that the segment is not passed down as a whole. An option to specify the survival probability of homozygote carriers of a deleterious allele or regions was also added. An interesting application of the simulation functions is to estimate

transmission probabilities for alleles or segments from ancestors to descendants, which can be computed quickly for simple relationships but can be difficult to obtain for complex genealogical structures.

Datasets and simulations

Two genealogical datasets are provided with the package, one is a fictive example of genealogical data and the other is a true genealogical dataset from the province of Quebec. The fictive example is included in the package for educational purposes. Users can explore the functions using this example, which is not too big but still has interesting features. In order to illustrate what can be accomplished with the GENLIB package, we performed a complete description of the ascending genealogies available for the 140 individuals in the Quebec sample. These individuals come from various regions of Quebec and this information is also provided as a separate dataset within GENLIB. This sample has already been investigated to demonstrate the presence of population structure in genetic data, supported by genealogical information (Roy-Gagnon *et al.* 2011). We then performed a simulation study to estimate the probability of sharing identical-by-descent (IBD) alleles and chromosomal segments of different lengths. Example source code for data analysis is provided in Supplementary Text S1 (p.102).

For the Quebec sample, all participants gave informed written consent, provided family information necessary to reconstruct their genealogy using the BALSAC database, and agreed to the public distribution of their genealogy in coded form. The study was approved by the CHU Sainte-Justine Ethics Committee (reference number: 2684).

Results

Description of genealogical data using GENLIB

The Quebec sample provided with GENLIB includes 140 individuals (probands) sampled from seven regional or ethno-cultural populations. The regions represented are Montreal, Quebec City area, Saguenay, North Shore and Gaspesia in which we distinguish 3 different populations (French Canadians, Acadians and Loyalists). The whole genealogical corpus includes 41,523 individuals (20,773 males and 20,750 females). We identified 21,230 nuclear families, including 5,994 full-sibships of sizes varying from 2 to 14 individuals. These are ascending genealogies, i.e., they are reconstructed for each of the 140 probands separately by identifying his/her ancestors and then linked together. Hence, the number and sizes of sibships are underestimated.

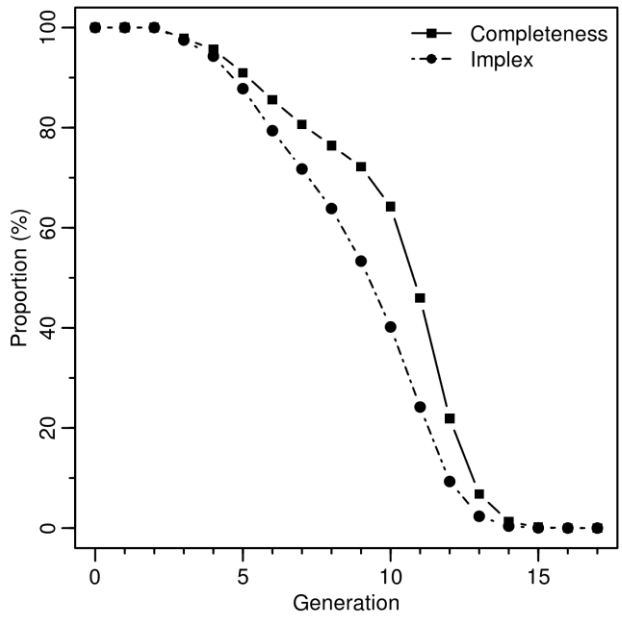

Figure 3.1 Completeness and implex indices for the Quebec genealogical corpus

Including the initial generation (generation 0), this corpus is 18 generations deep. We identified 7,399 founders, i.e., individuals without parents in the BALSAC database, among all individuals in the genealogy. A few founders are in the second generation, meaning that some lineages stop after grandparents. However, more than 90% of the founders are in the 7th generation or higher. The mean genealogical depth is 9.4, indicating that on average genealogical lineages are 9.4 generations deep. Computed values of completeness and implex indices are shown on Figure 3.1 (p.88). From the initial generation to the second back in time we observe that the whole genealogical corpus is known (completeness equals 100%) and its size grows exponentially from a generation to another (implex also equals 100%). From the third generation, the completeness and implex indices drop below 100% because pedigree collapsing starts and some information is missing. This means that 1) some lineages stop at the second generation, as noted earlier by the presence of founders in that generation, and 2) at least one individual has related ancestors in the first 3 generations, as indicated by an implex value slightly smaller than the completeness value. We also note that the implex index decreases faster than the completeness and that the gap between these two indices is the largest at the 10th generation.

Figure 3.2 (p.91) presents the relationship between the cumulative proportion of the genetic contribution of founders and the cumulative proportion of contributing founders. We observe that the genetic contribution of founders varies across populations with values quite far from the linear relationship resulting from the theoretical (unrealistic) uniform genetic contribution that would be obtained if every founder had an equal contribution to the gene pool. The sample from the Acadian sub-population shows the most uneven contribution of founders as 2.3% (n=49) of their founders contribute to 50% of the gene pool. In comparison, the sample from Montreal has 14.4% (n=794) of its founders contributing to 50% of the genetic pool. The larger contribution of some founders is explained by their differential demographic history, as well as the one from their descendants, in addition to their time of arrival in Quebec. Indeed it was previously observed that kinship coefficients vary among these populations (see Figure 4 from (Roy-Gagnon *et al.* 2011)). The highest inbreeding coefficient among all individuals in the genealogy (0.17) is observed for a proband from the Loyalist population. The parents of this individual are double first cousins, his grand-parents are double second cousins and his

great-grand-parents are also first cousins (Figure 3.5 p.100). Within the whole genealogy, 3.7% (n=1549) of people have inbreeding coefficients greater than 1/64 corresponding to parents that are second cousins.

The founders with the highest genetic contribution are not the ones with the highest coverage, i.e., linked to the maximum number of probands, because the genetic contribution is higher when an ancestor is closer while coverage increases as an individual is more distant. The high coverage individuals are all founders except one (who is the son of two of these high coverage founders) and each is linked to the same 121 probands out of all 140 probands in the data. The probands not covered by these prolific ancestors are mainly descendants of Loyalists, who arrived in Quebec after the American War of Independence, had little intermarriage with the surrounding French Catholic population and had less complete genealogical data. For distant ancestors, the genetic contribution is correlated with the number of occurrences of an ancestor within an individual's genealogy, in other words how many different, but not necessarily mutually exclusive, paths link an ancestor to a descendant. The number of occurrences is maximal for a proband from Saguenay and some of its ancestors. Four founders and two of their children are linked 176 times trough different paths to this proband. These four founders also have the maximal genetic contribution in this genealogical corpus. All probands with ancestors occurring more than 60 times in their genealogies come from Saguenay and North Shore. Moreover, even if probands from the Acadian population have higher kinship values than those from the North Shore and Saguenay (Roy-Gagnon *et al.* 2011), we observe that North Shore and Saguenay are two populations showing an accumulation of distant common ancestors with high numbers of occurrences in the genealogy.

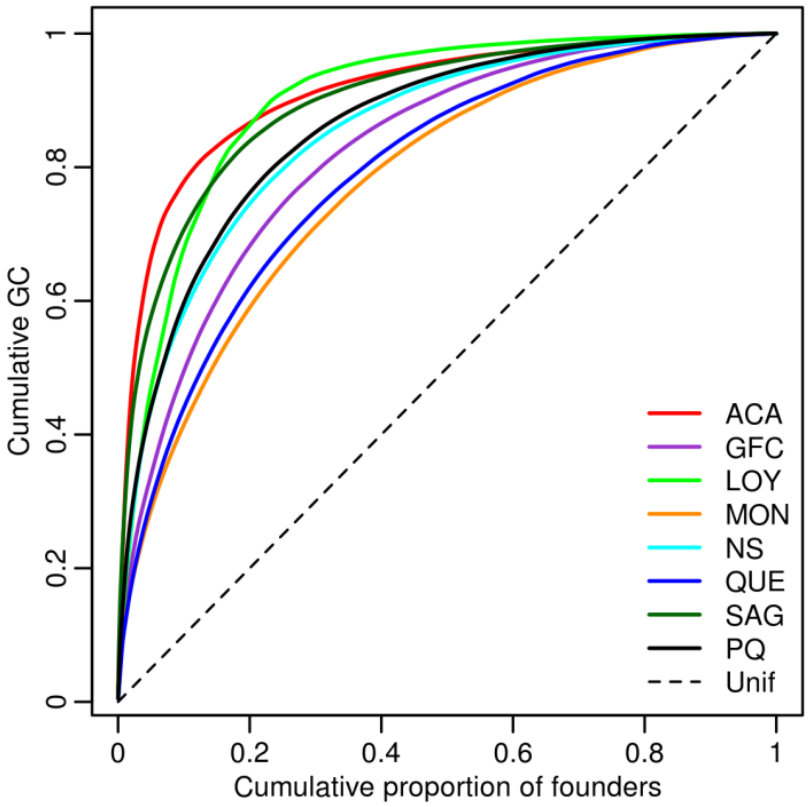

Figure 3.2 Cumulative genetic contribution of founders for each population

Plot of the cumulative distribution of genetic contributions of founders for each population in relation to the cumulative proportion of contributing founders, sorted in decreasing order of their genetic contribution. The dashed line presents the hypothetical situation in which all founders of a population contribute equally to the gene pool. ACA Acadians, GFC Gaspesian French Canadians, LOY Loyalists, MON Montreal, NS North Shore, QUE Quebec City area, SAG Saguenay, PQ Whole sample from the Province of Quebec, Unif Uniform distribution

Gene-dropping simulations using GENLIB

We performed gene-dropping simulations to evaluate the odds of IBD sharing for two pairs of distantly related individuals. We selected individuals that had similar kinship coefficients but different patterns of genealogical links (e.g., differences in the number of common ancestors and distances to these ancestors), so that patterns of IBD sharing are expected to differ between the pairs. For each pair of individuals, we identified all common ancestors and used these common ancestors as starting point for allele dropping simulations. We performed 10,000,000 simulations for each common ancestor and calculated the probability of IBD sharing as the proportion of simulations where at least one allele was shared. One pair of individuals is from the Saguenay population with a kinship coefficient of 0.0051 and shares 1628 common ancestors. The other pair of individuals is from the Acadian population with a kinship coefficient of 0.0044 and shares 261 common ancestors. Realized IBD sharing and genealogical measures have previously been compared between samples from the Saguenay and Acadian populations, indicating overall higher levels of IBD sharing in the Acadians in agreement with the observed smaller number of closer common ancestors (Gauvin *et al.* 2014). Our goal here was to compare IBD sharing at distant kinship levels that are similar but arise from very different genealogical links, such as those present in the Acadian and Saguenay genealogical structures.

One advantage of gene-dropping simulations is that they give a more accurate picture of an ancestor's impact on the genome of two individuals than the genetic contribution does. As can be seen on Figure 3.3 (p.93), estimated probabilities of sharing one allele IBD are either perfectly correlated with the product of the two genetic contributions, either greater, depending on the ancestor from whom this allele was inherited (Supplementary Figure 3.6 (p.101) illustrates the different types of ancestors). The odds of sharing an allele IBD depend on the number of meioses separating two individuals through a particular common ancestor and also on all the possible paths linking those two through this common ancestor. When we consider MRCAs, the product of an ancestor’s genetic contributions to the two probands is equal to the probability that the probands share one allele IBD because the paths from a proband to the MRCA and from the MRCA to the other proband are completely distinct.

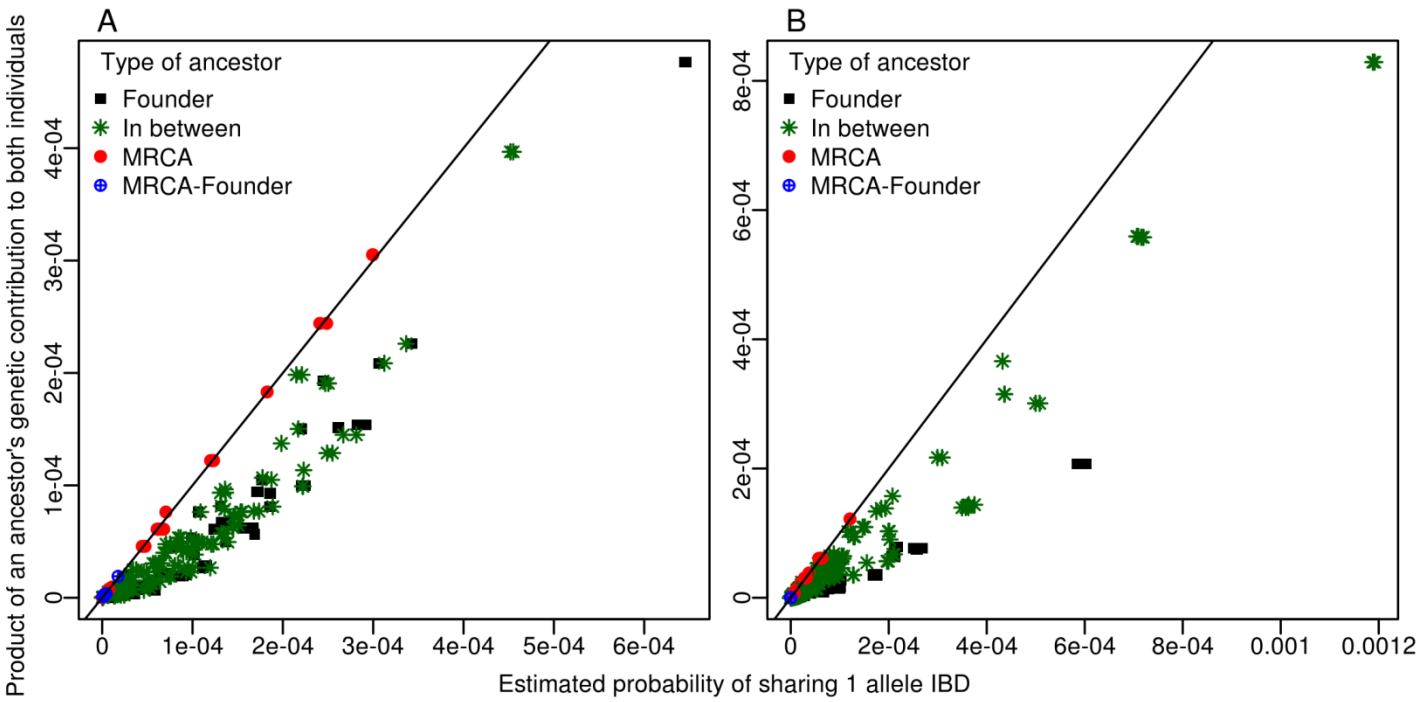

Figure 3.3 Estimated probabilities of sharing one allele IBD versus ancestors' genetic contributions

Plots of estimated probabilities of sharing one allele identical-by-descent (IBD) from a specific ancestor relative to the product of the genetic contributions of that ancestor to each of the two individuals from A) the Acadian population and B) the Saguenay population. Probabilities that the two individuals share one allele IBD were estimated from 10,000,000 gene-dropping simulations for each shared ancestor. Ancestors are divided into four categories depending on whether they are founders, most recent common ancestors (MRCA), both (MRCA-Founder) or neither of the two (In between). The black line is the identity line, i.e. y=x.

Otherwise, for non-MRCA ancestors (either founders or those in between founders and MRCA), the product of genetic contributions from that ancestor to both probands underestimates the probability of sharing an allele IBD. The more distantly linked two individuals are, the less likely they are to share an allele IBD and their variance for the number of IBD sharing occurrences across simulation replicates is also lower (Donnelly 1983; Hill 1993).

We extended the simulation functionalities of GENLIB to simulate the transmission of chromosomal segments for each pair of individuals. Using IBD sharing previously inferred from genome-wide genotype data available on the 140 individuals (Gauvin *et al.* 2014), we selected multiple segments shared IBD by the two pairs of individuals described above. We used these observed segments to select the lengths of the segments to be simulated. This second set of simulations includes a probability that the segment undergoes recombination. For all segments we considered recombination rates based on sex-specific recombination maps (Kong *et al.* 2010) (see Table 3.3). We performed 100,000,000 simulations for each common ancestor. We were interested in the odds that the segment is transmitted intact from a common ancestor to both members of the pair, resulting in a segment shared IBD.

Table 3.3 Selected segments shared IBD by two pairs of individuals

Individual IDs	Chr	Length (cM)	Length male recombination map (cM)	Length female recombination map (cM)
408868, 409033 Acadian	1	3.5416	3.0829	4.2469
	5	7.3682	5.0353	9.0442
	8	10.0267	4.0278	14.9491
302710, 302711 Saguenay	3	3.3035	2.2441	4.4612
	7	5.8712	4.0323	8.4025
	1	12.6098	7.1124	17.3889

The results of the segment-dropping simulations were concordant with expectations. Indeed we observed that the probability of sharing a segment IBD decreases compared to the probability of sharing only one allele IBD with the increasing length of segments (Figure 3.4). Linear regressions for each segment length show how estimated probability of sharing a segment IBD decreases with an increasing recombination rate. Our ability to discriminate

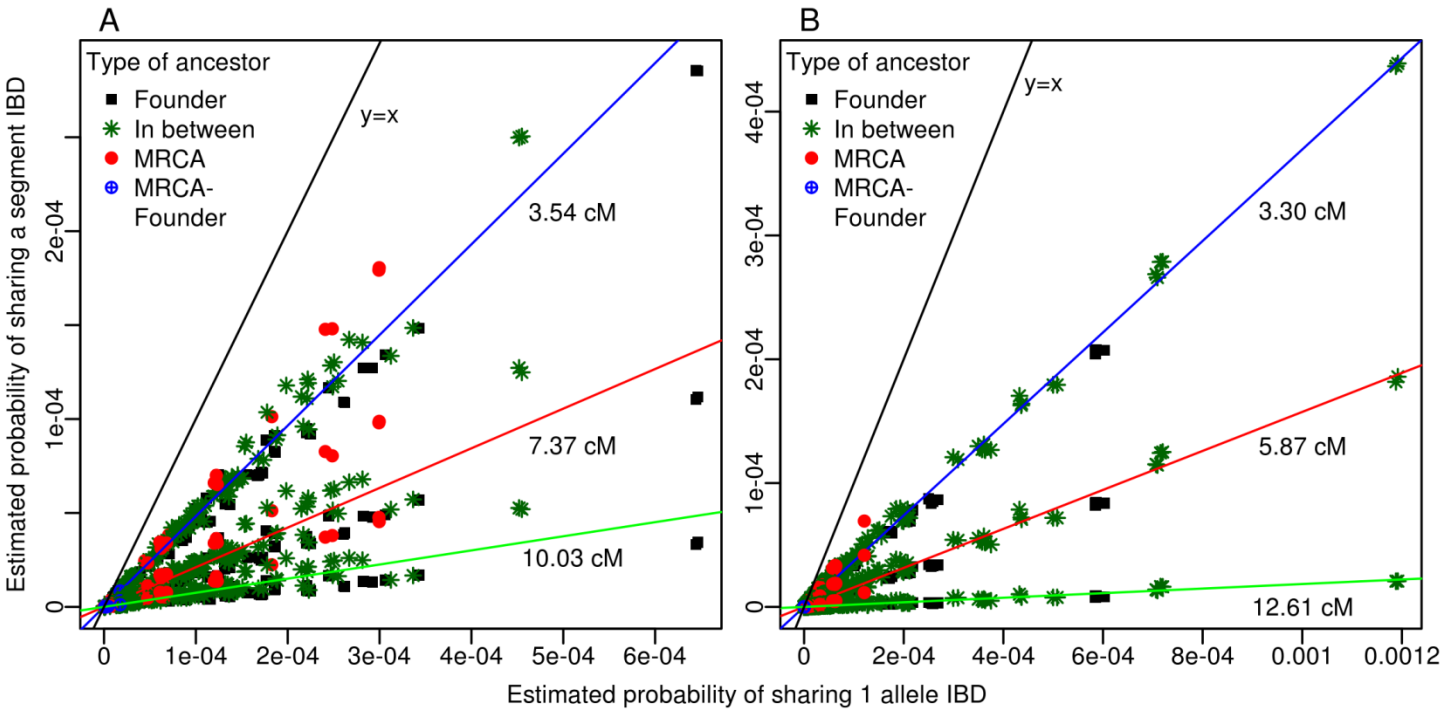

Figure 3.4 Estimated probabilities of IBD sharing for a segment versus one allele

Plots of estimated probabilities of sharing one segment identical-by-descent (IBD) from a specific ancestor relative to the estimated probabilities to share one allele IBD from that same ancestor for a pair of individuals from A) the Acadian population and B) the Saguenay population. Probabilities that the two individuals share one allele or one segment IBD were estimated from 10,000,000 (for allele sharing) or 100,000,000 (for segment sharing) simulations for each common ancestor. Ancestors are divided in four categories depending on whether they are founders, most recent common ancestors (MRCA), both (MRCA-Founder) or neither of the two (In between). The solid black line is the identity line and colored lines are simple regression lines between IBD sharing of a segment and IBD sharing of an allele. Three different segment lengths are considered and shown for A: 3.30, 5.87 and 12.61 cM and for B 3.54, 7.37 and 10.03 cM (Table 3.3).

which ancestors could have contributed IBD segments is also facilitated with longer segments. Using a threshold on the probability of IBD sharing of 1.00E-08, corresponding to the 10th percentile of the overall probability distribution, we found that we could eliminate proportionally more potential ancestors unlikely to have transmitted a segment for the pair of individuals from Saguenay compared to the Acadians for similar segment lengths. For the smallest segment (~3.4 cM) we can exclude, respectively for the two Acadians and the two Saguenay individuals, 0 (0.0%) and 77 (4.7%) ancestors, for the midsize (~6.6 cM), 6 (2.3%) and 154 (9.5%) ancestors and for the longest segment size (~11.3 cM), 20 (7.7%) and 528 (32.4%) ancestors. As expected, segment length plays an important role since the longer the segment is, the higher the recombination chances are. Another important factor is the length of inheritance paths. Each time a segment is transmitted, it is subject to recombination so the longer a path between two individuals through an ancestor is, the higher the odds of recombination are.

Discussion

In this paper, we presented GENLIB, an R package for the analysis of complex genealogical data. GENLIB can handle large extended human pedigrees, extended genealogical data from human founder populations or extended animal pedigrees. We were able to read and describe a fictive pedigree with a size over 10,000,000 and a depth of 58 generations. Some memory limits may be encountered, depending on the computer used, when trying to perform calculations (e.g., kinship) for a large number of individuals at the same time in complex genealogies, but these issues can easily be fixed by performing calculations in batches instead.

GENLIB is designed to be easily accessible and used by researchers familiar or not with genealogical information. The only requirement is a basic knowledge of the R environment. GENLIB includes functions to describe genealogies, to compute different genealogical indices and to perform simulations based on genealogical relationships. Given the flexibility of the R environment, users can also create their own functions to further describe, summarize and present (e.g., graphs) GENLIB results.

We used genealogies from the Quebec founder population to illustrate the utility of GENLIB to describe and analyze genealogical data. In addition, we performed simulations to track the transmission of alleles or chromosomal segments through the genealogy. Allele- or segment-dropping simulations that consider the complete genealogical structure can help to predict the risk of sharing disease mutations introduced by founders, or to identify ancestral sources of variation in a trait of interest (Vézina *et al.* 2005a; Chetaille *et al.* 2014). In our simulations, we highlighted the fact that specific population characteristics, such as the number of shared ancestors and the distance to these shared ancestors, can help discriminate among founders more or less likely to have transmitted segments IBD. The fact that individuals from Saguenay share more common ancestors that are more distant on average compared to Acadians explains, in part, why, for similar kinship values, we can observe quite different IBD segment distributions. Gene-dropping simulations can also be used to predict the risk of future loss of genes contributed by the various founders. In addition, distributions of IBD probabilities for any individuals can be generated through simulations.

We plan to continue to improve existing functions, for example by adding more flexible options for individual fitness in the simulations. We also plan to extend the range of functions available. Examples of planned additions include implementing more data importation and graphical functionalities, incorporating information about probands and ancestors, such as birth places and dates, implementing a function to flag unrelated components (subpedigrees) within a kindred, and implementing statistical tests to compare kinship and inbreeding distributions between different groups. Another planned addition is to include an option in the segment-dropping simulations to recover the length of transmitted segments instead of only its status (fully transmitted or not). With its flexibility, GENLIB provides an important contribution towards an improved and optimal way to combine information coming from genealogies and genetic material. These two sources of data are complementary and using them together could improve genetic data analysis in founder populations, for example in the context of estimating relatedness, imputing genomic data or haplotyping (Kong *et al.* 2008; Roy-Gagnon *et al.* 2011; Speed and Balding 2015).

Conclusions

The GENLIB R package was developed to facilitate a broader and more extensive use of available genealogical data from founder populations across the world. The package provides a user friendly and flexible environment to analyze genealogical data, allowing a more efficient and easier integration of different types of data and analytical methods and making it ideal for further developments. Using GENLIB simply requires a minimal knowledge of R and provides many functions to manage, describe and analyze genealogical data. Our description of a Quebec genealogical sample and our simulation study illustrated the use of GENLIB and further highlighted regional differences present in the population. We also provided insight on factors influencing the resulting IBD sharing in founder populations. GENLIB is an improvement over existing software in that it provides gene- and segment-dropping simulations and other new functionalities and in that it can be used within R, which is a computing environment familiar to many statistical geneticists. Hence, it is a highly valuable tool to researchers studying extended genealogies.

Supplementary Figures and Tables

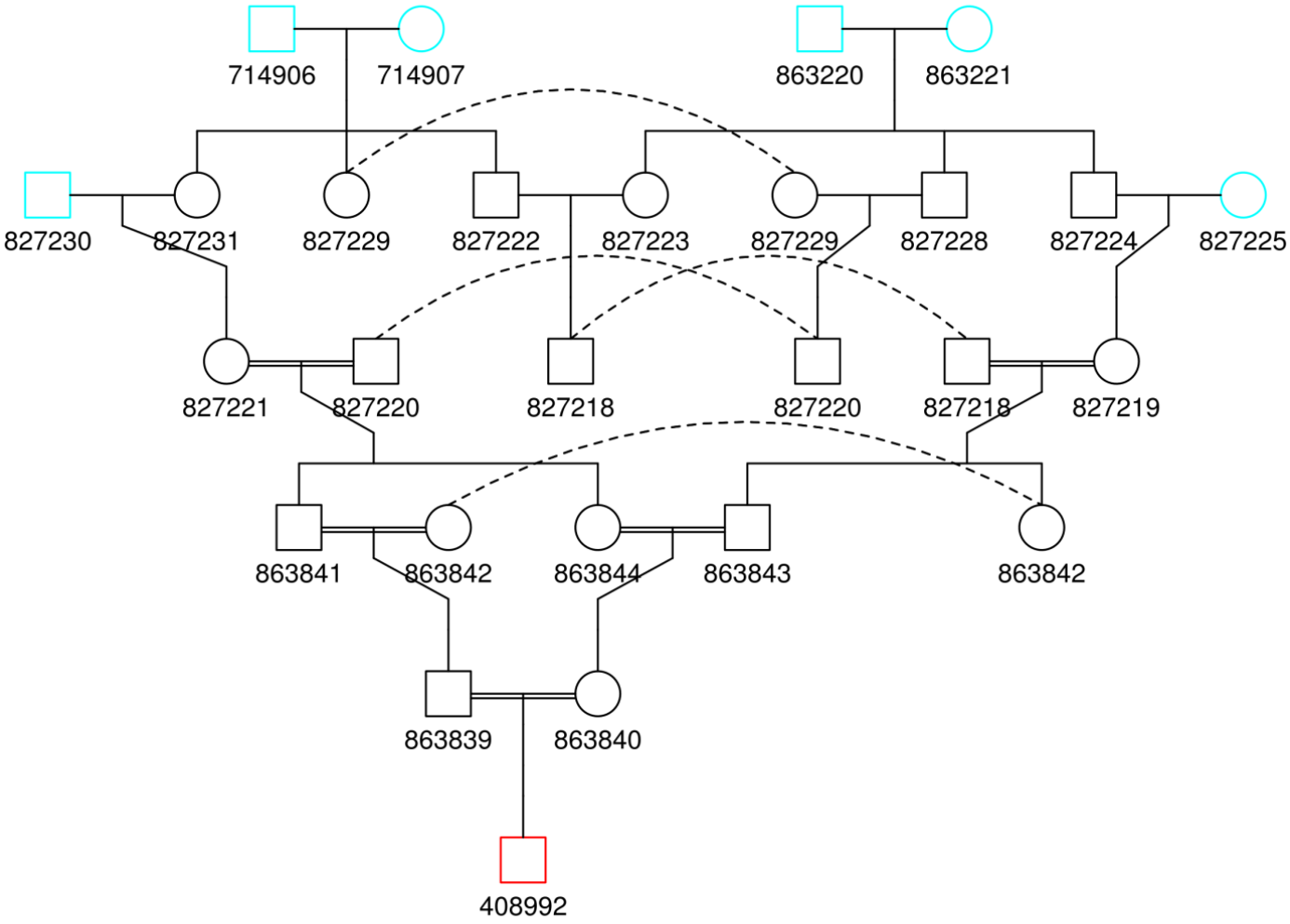

Figure 3.5 Genealogy of a highly inbred individual

Genealogical tree for one individual from the Loyalist population. Lineages are cut as soon as unrelated individuals are found.

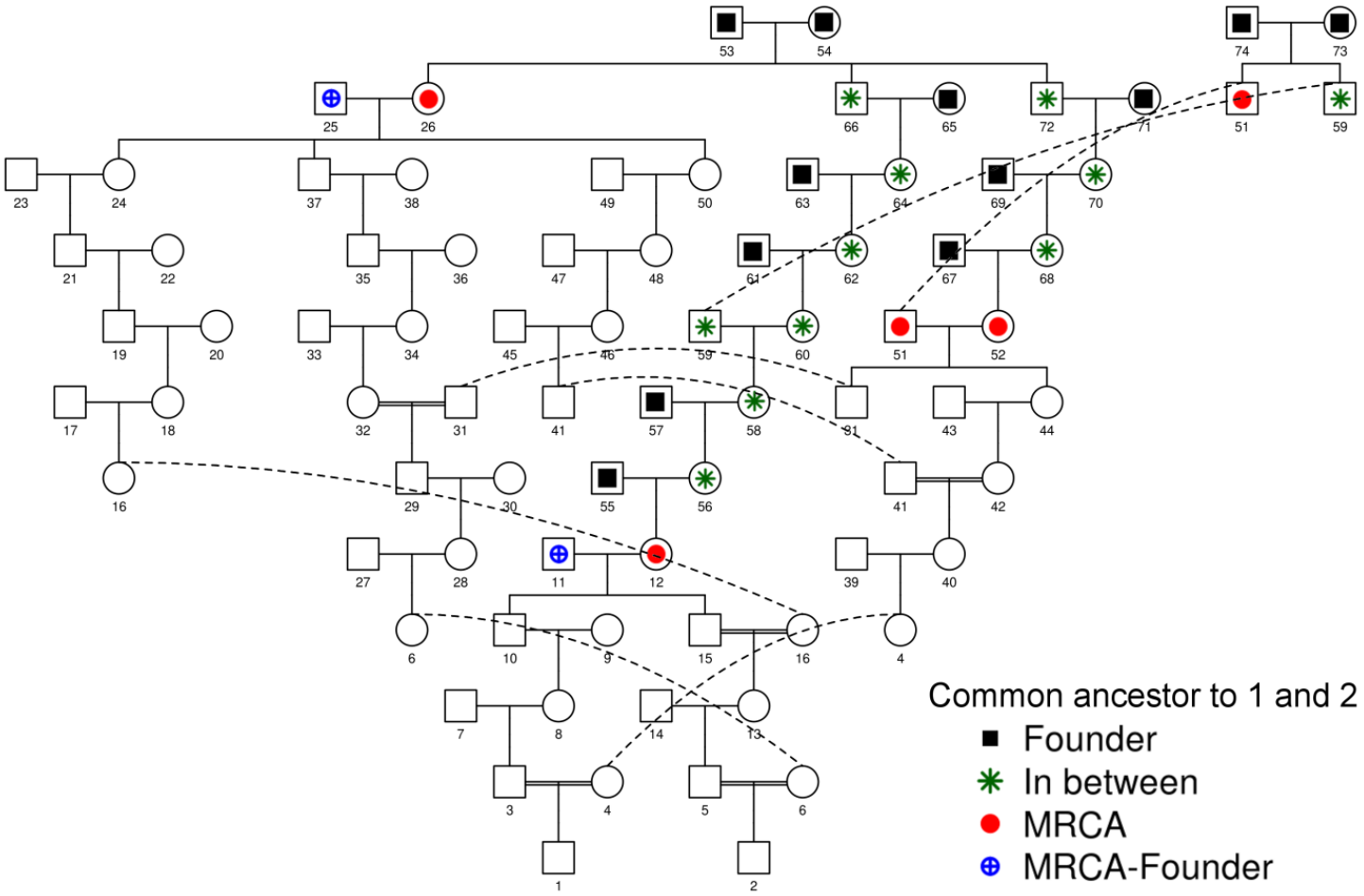

Figure 3.6 Genealogical example showing different types of common ancestors

Two probands can share either 1) most recent common ancestors (MRCA), 2) MRCA which are also founders (MRCA-Founder), 3) ancestors between MRCA and founders called “In between” or 4) founders.

Supplementary Text 1

Example source code for the description and analysis of the genealogical data presented.

```
##########################
##########################
##### Load the library
library("GENLIB")

##### Preview data available into the GENLIB package, type "q" to quit
#data(package="GENLIB")

##### Load and preview the genealogical dataset from Quebec
data(genea140, package="GENLIB")
head(genea140)

##### Load and preview the population of origin of the 140 probands from 'genea140'
data(pop140, package="GENLIB")
head(pop140)

##### Numbers of individuals per population
table(pop140[,2])

##### To get help on a function, type '?' + function and type "q" to quit
#?gen.genealogy

##########################
##########################
##### Create the genealogical object
gen140<-gen.genealogy(genea140)
##### Print basic information about the genealogy
gen140

##### Alternatively
gen.noind(gen140)
gen.nomen(gen140)
```

```
gen.nowomen(gen140)

############################
############################  Warning : this part can take a few minutes (~20)
##### Siblings and sibships size
sibshipsize<-NULL
##### If individuals have no parents, then we can not known if they are siblings
##### Those individuals are removed from the list of potential siblings
list_individuals<-genea140[-which(genea140[,2]==genea140[,3] & genea140[,2]==0),1]
i=1

while (i<=length(list_individuals)){

        ##### Test if an individual has full siblings
        sibs<-gen.sibship(gen140,list_individuals[i], halfSibling=FALSE)
        # Add the size of this sibship to the list
        sibshipsize<-c(sibshipsize,length(sibs)+1)

                # Remove siblings from the list of individuals to test because we already
                # know they do have siblings and their family size is recorded
                if(length(sibs)>0){
                list_individuals<-list_individuals[-match(sibs,list_individuals,0)]
                }

        i<-i+1
}

table(sibshipsize)

##### 21230 Nuclear families
sum(table(sibshipsize))
##### 5994 Sibships of size >= 2
sum(table(sibshipsize)[-1])

############################
############################
##### Number of generations
```

```
gen.depth(gen140)

##### Identify all founders in the genealogy
founders<-gen.founder(gen140)
length(founders)
##### Identify all half-founders in the genealogy
half_founders<-gen.half.founder(gen140)
length(half_founders)

##### Generation to which the founders appear first (minimal generation)
minfound<-gen.min(gen140, individuals=founders)
table(minfound)
##### More than 90% of the founders are found in the 7th generation or later
table(cut(minfound, breaks=c(1,6,14)))/length(founders)

##### Mean genealogical depth
gen.entropy(gen140)

##########################
##########################
##### Completeness and implex indices
completeness_values<-gen.completeness(gen140)
implex_values<-gen.implex(gen140)

completeness_values
implex_values

############################
############################
##### Genetic contribution
##### Vector of populations' names
names_pop<-names(table(pop140[,2]))

##### List of founders from each population
list_fond_per_pop<-NULL
for(i in 1:length(names_pop))
{
        gen_1pop<-gen.branching(gen140,
```

```
        pro=pop140[which(pop140[,2]==names_pop[i]),1])
        founder_1pop<-gen.founder(gen_1pop)
        list_fond_per_pop<-c(list_fond_per_pop, list(founder_1pop))
}

names(list_fond_per_pop)<-paste(names_pop, "_founders", sep="")
str(list_fond_per_pop)

##### Genetic contribution of all founders to each proband
gen_contrib<-gen.gc(gen140, ancestors=founders)

##### List of genetic contributions, sorted in decreasing order, of founders of each population
to all probands
GC_per_pop<-NULL
for(i in 1:7)
{
        GC_1pop<-
        colSums(gen_contrib[match(pop140[which(pop140[,2]==names_pop[i]),1],gen.pro(ge
        n140)),match(list_fond_per_pop[[i]], founders)])
        GC_per_pop<-c(GC_per_pop, list(sort(GC_1pop, decreasing=TRUE)))
}
names(GC_per_pop)<-paste("GC_", names_pop, sep="")
str(GC_per_pop)

quantile(cumsum(x=GC_per_pop$"GC_Gaspesia-
Acadian"),probs=c(0.0231))/sum(GC_per_pop$"GC_Gaspesia-Acadian")
quantile(cumsum(x=GC_per_pop$"GC_Montreal"),probs=c(0.1433))/sum(GC_per_pop$"GC
_Montreal")

############################
############################
##### Inbreeding coefficients for all the individuals included in the genealogy
inbreeding_coeff<-gen.f(gen140,pro=genea140[,1] )

##### Maximum inbreeding coefficient, ID of the individual and its origin
max_inbred<-max(inbreeding_coeff)
max_inbred
genea140[which(inbreeding_coeff==max_inbred),1]
```

```
pop140[match(genea140[which(inbreeding_coeff==max_inbred),1], pop140[,1]),2]

##### Number of individuals with inbreeding coefficient in the range 0, 1/64 (parents 2nd cousins), 1/16 (parents cousins) and 1/4
table(cut(inbreeding_coeff, c(0,1/64,1/16,1/4), right=FALSE))

############################
############################
##### Coverage, maximal value and ancestors' ID with the highest values
coverage<-gen.rec(genea140)
max_cov<-max(coverage)
max_cov
genea140[which(coverage==max_cov),1]

##### Coverage of the individuals with the highest genetic contribution
coverage[names(GC_per_pop[[1]][1:2]),]

##### Occurence, maximal value and individuals with highest values
occurence<-gen.occ(genea140)
max_occ<-max(occurence)
max_occ
who_max_occ<-which(occurence==max_occ, arr.ind=TRUE) ##Same proband having the ancestors with highest occurences
occurence[who_max_occ[,1], who_max_occ[,2]]

##### Origin of probands not covered by the ancestors with the highest coverage
table(pop140[match( names(which(colSums(occurence[which(coverage==max_cov),])==0) ), pop140[,1]),2])

##### Origin of probands covered more than 60 times by ancestors
table(pop140[match( dimnames(occurence)[[2]][unique(which(occurence>60, arr.ind=TRUE)[,2])], pop140[,1]),2])

############################
############################
##### Gene- and segment-dropping simulations
```

```
##### Using Saguenay individuals as the pair of interest
pop140[c(5,19),]

##### Set the number of simulations
no_simu<-100000

##### Simulation for one gene coming from founder #18322
simuRes1<-gen.simuSample(gen140,pro=c(302710, 302711), ancestors=18322,
stateAncestors=1, no_simu)
##### Number of genes transmitted at each individual
table(simuRes1)
##### Number of genes transmitted to both individual at each simulation
table(colSums(simuRes1))
##### 2 or more copies of the gene can be transmitted, but it does not mean our 2 individuals
share one identical-by-descent
##### (same gene transmitted from one ancestor to both individuals) because only one could
have received 2 copies and the other 0
##### Estimated probability that the gene is transmitted twice to one of the 2 individuals
sum(simuRes1[1,]==2 |simuRes1[2,]==2 )/no_simu
##### Estimated probability that an gene is transmitted identical-by-descent (at least one copy
each)
sum(simuRes1[1,]>=1 & simuRes1[2,]>=1 )/no_simu

##### Simulation for one segment ~5.87 cM (having recombination rate of 0.040323 for male
and 0.084025 for female)
proba_segment<-gen.simuProb(gen140, pro=c(302710, 302711),statePro=c(3,3),
ancestors=18322, stateAncestors=1, 100000, probRecomb = c(0.040323, 0.084025))
##### Returns the probability that the joint event happens, meaning that at least one segment
is transmitted to each proband,
##### the probability that each proband receive at least one segment and the probability that 0,
1 and 2 probands inherit at least one segment.
proba_segment
```

Chapter 4:
Distant inbreeding among French Canadians and associations with health-related traits

Héloïse Gauvin, Jean-Christophe Grenier, Vanessa Bruat, Elias Gbeha, Élodie Hip-Ki, Marie-Hélène Roy-Gagnon and Philip Awadalla

Reference: Gauvin H, Grenier J-C, Bruat V, Gbeha E, Hip-Ki E, Roy-Gagnon M-H, Awadalla P. Distant inbreeding among French Canadians and associations with health-related traits. *To be submitted*

Authors' contribution

In this paper, my contribution is:

- The literature review on inbreeding depression;
- The bioinformatic and statistical analysis of the genetic data;
- The simulation study and its analysis;
- The design of the study with MHRG and PA;
- Writing of the paper.

Contributions of other authors are: JCG preprocessed the genomic data, performed quality control and provided bioinformatic support. VB provided bioinformatic support. EG and EHK processed samples for genotyping. PA provided samples. All authors reviewed and approved the manuscript.

Acknowledgements

We thank the CARTaGENE participants and team for data collection. We acknowledge financial support from Fonds de la Recherche en Santé du Québec (FRSQ), Génome Québec and the Canadian Partnership Against Cancer. HG is the recipient of a scholarship from the Réseau de Médecine Génétique Appliquée (RMGA).

Abstract

Inbreeding depression is a phenomenon causing a loss of fitness for offspring of related parents. This decrease in fitness could affect fertility or ability to survive. Different studies in human populations reported effects on physiological traits but the lack of consistent evidence for some associations challenges the direct impact of inbreeding on health. We used a group of participants (n=958) from the CARTaGENE population-based biobank with known genotypic information and numerous phenotypes to perform population genetic analyses and to investigate the inbreeding burden among French Canadians, who represent the majority of people recruited in this biobank. The French Canadian population is a founder population with a unique population structure and varying rate of distant consanguinity. Population genetic analyses were done in order to select only French Canadians from all the participants. We identified stretches of homozygous genotypes within French Canadian individuals (n=727) and we used them to assess each individual's genome-wide level of inbreeding, which corresponded, on average, to those of offspring from parents being third degree cousins. In our study we examined the correlation of different traits with inbreeding using simple linear regressions adjusted for age and sex. We replicated previous associations with height and body mass index, and we identified novel potential associations with white blood cell and neutrophil counts, which could potentially influence disease susceptibility through the efficiency of the immune system.

Introduction

Runs of homozygosity (ROHs) are long stretches of consecutives homozygous genotypes. Their distribution in an individual's genome depends on the population from which the individual comes from and more directly on the relationships shared by the individual's ancestors (McQuillan *et al.* 2008; Gusev *et al.* 2012). Indeed, the more closely related two individuals or parents are, the higher are the chances that their offspring will receive identical chromosomal segments from them. Hence, to infer the level of inbreeding in a population one option is to track these homozygous segments within individuals. Studies have shown that ROHs are common in outbred populations (Gibson *et al.* 2006; Li *et al.* 2006; Pemberton *et al.* 2012), and these segments have generated great interest in part because of their potential impact on health (Ku *et al.* 2011). A recent study showed that homozygosity stretches generate genomes having an increased burden of deleterious recessive variants (Szpiech *et al.* 2013).

Increased homozygosity for recessive detrimental variants is one explanation for the inbreeding depression phenomenon that causes offspring of related individuals to show reduced fitness, e.g. lower fertility and survival rate, and in some cases major abnormalities (Charlesworth and Willis 2009). The other explanation for inbreeding depression rests on the potential advantage of being heterozygous, such as at HLA loci (Black and Hedrick 1997). In the case of HLA, balancing selection maintains heterozygous loci at frequencies higher than expected, probably because being heterozygous at that loci confers an advantage over the homozygous status.

Detrimental effects of inbreeding depression have long been known and are well documented for plants and animals, in part because they lend themselves more readily to experimental designs. In humans, inbreeding depression has been reported for multiple complex traits such as height (McQuillan *et al.* 2012), blood pressure (Rudan *et al.* 2003b, 2006; Campbell *et al.* 2007), cholesterol (Campbell *et al.* 2007; Panoutsopoulou *et al.* 2014), red blood cells (Panoutsopoulou *et al.* 2014) and haemoglobin (Panoutsopoulou *et al.* 2014), and for diseases such as schizophrenia (Lencz *et al.* 2007; Keller *et al.* 2012), Alzheimer (Vézina *et al.* 1999;

Farrer *et al.* 2003; Nalls *et al.* 2009) and coronary heart disease (Rudan *et al.* 2003a; Ismail *et al.* 2004).

One potential reason for inbreeding depression affecting some traits more than others is the heritability of these traits. The strongest evidence for inbreeding depression in a human trait is for height, which is highly heritable with estimates around 80% (Pilia *et al.* 2006; Visscher *et al.* 2006; Perola *et al.* 2007). Other traits such as high density lipoprotein (HDL) and low density lipoprotein (LDL) cholesterol have heritability estimates between 22 % and 63% (Abney *et al.* 2001; Pilia *et al.* 2006; Macgregor *et al.* 2010), which make them perhaps less prone to inbreeding depression.

Our study involved French Canadian (FC) inhabitants of the relatively recently founded province of Quebec in Canada. The initial French founders settled 400 years ago in the St-Lawrence river valley. A century and a half later, French immigration practically ceased with the British Conquest. The population then grew steadily and in relative isolation because of language and religious barriers. These barriers amplified the founder effect, including the colonisation of new remote and isolated regions within Quebec, as the population increased. Although the province welcomed immigrants during the 19th and 20th century, they had limited impact on the FC genetic pool (Vézina *et al.* 2005b; Bherer *et al.* 2011). Today up to 80% of the 8 million inhabitants of Quebec province are of FC descent. Many aspects of the FC founder population have been studied encouraged by the presence of extensive genealogies capturing the unique demographic history of this population (Roy-Gagnon *et al.* 2011; Moreau *et al.* 2011). Regarding the population history, two recent studies investigated how it may have influenced the genetic fitness and the accumulation of mutations in the FC population compared to the progenitor French population as well as other worldwide populations (Casals *et al.* 2013; Hussin *et al.* 2015).

Using genotypic and phenotypic information of deeply phenotyped participants of the CARTaGENE Project in Quebec, we performed population genetic analyses and investigated the inbreeding burden among FC. Our genomic data showed very detailed population structure in the modern Quebec population. The identification of FC was aided by techniques that take

advantage of the haplotypic structure. Estimated inbreeding levels for French Canadians corresponded, on average, to those of offspring from third degree parents. Our results corroborated previous findings about inbreeding depression affecting height, however not other phenotypes previously identified. Still we identified novel associations between inbreeding and white blood cells (WBC) count and one of its cell subtypes, neutrophils, and also body mass index (BMI). Although, as far as we know, we are the first study to report the association with WBC; in animal studies increased inbreeding was associated with an increase in infectious diseases (Ilmonen *et al.* 2008; Murray *et al.* 2013; Hoffman *et al.* 2014). If inbreeding is indeed reducing the number of WBC, and therefore reducing the efficiency of the immune system, this could explain an increased susceptibility to infections of all kinds for more inbred individuals, as observed in other studies.

Materials and methods

Participants and phenotyping

The CARTaGENE project is a prospective population-based cohort in which participants were recruited on the basis of their age (between 40 and 69 years old) and place of residence, the province of Quebec in Canada. As of today, the cohort includes over 40 000 individuals deeply phenotyped with different questionnaires on health, biospecimen collected and physical measurements taken during a visit to an assessment site (Awadalla *et al.* 2013; CARTaGENE 2015). For each individual one of the blood samples collected was sent to a central laboratory for haematological and biochemical tests (complete blood count assay). For the present study, we used a set of 958 individuals sampled in three regions: Montreal, Quebec City and Saguenay. The individuals were previously selected for a study on environmental variation of gene expression profiles (Fave *et al.* 2015). The selection was conducted in order to get an equal representation of ages and sexes and a range of arterial stiffness values. For our study we selected all traits previously linked to inbreeding depression, including anthropometric, haematological and biochemical traits and traits relating to lung functions and the circulatory system. Details about the different physiological measurements are available in Awadalla *et al.* (2013). In order to compare our sample of FC individuals to the whole CARTaGENE cohort, we computed descriptive statistics about all phenotypes. For the interpretation of the results we also looked at correlation between the phenotypes.

Most participants (n=902) provided information on their ancestral origin by means of information on their 4 grandparents' country of birth. Individuals with at least 3 grandparents born in the same country are reported as coming from this country, otherwise they are considered admixed or with an unknown provenance (DNK: do not know) when more than half of the information on grandparents is missing (n=15). Ancestral origin was further investigated at the genetic level. All Canadian individuals were considered as FC unless genetic information disagrees with that assumption. CHU Sainte-Justine ethics committee

approved the study protocol and all participants included in the study gave their informed consent.

Genotyping

All individuals were genotyped on the Illumina Omni2.5M array. A total of 2,013,457 autosomal SNPs were obtained after quality control (HWE <0.001, SNP missing rate <0.10). Quality control filters were applied using PLINK v1.08 (Purcell *et al.* 2007). All genomic positions are according to NCBI build 37.

Population genetics

The study is focused solely on individuals with a FC ancestry therefore we have to identify them among all samples from Quebec. As a first step traditional principal component analysis was performed on the genotypes (PCAgeno) with EIGENSOFT v6.0.1 (Patterson *et al.* 2006). For PCAgeno we used SNPs with MAF >5% and pruned for LD (pairwise r^2<0.2 and 50 SNPs window shifting every 5 SNPs), yielding 146,689 SNPs.

In order to unveil finer scale patterns of population structure, i.e. differences among individuals with European ancestry and individuals having a FC ancestry, we also used ChromoPainter v0.04 (Lawson *et al.* 2012). We used all SNPs data apart from singletons, yielding 1,908,336 SNPs. Singletons were removed as they are non-informative for phasing and contribute to computation burden for haplotypes sharing inference performed with ChromoPainter. Genotypic data was phased with Shape-It v2.r644 (Delaneau *et al.* 2012) using the HapMap genetic maps (International HapMap Consortium *et al.* 2007). Coancestry matrices were obtained from ChromoPainter with parameters estimation steps performed with 10 iterations on 4 chromosomes only. The ChromoPainter method performs a reconstruction of every individual genome using chunks of DNA from the other individuals and reports matrices of the number and length of those shared chunks. We used the chunk count matrix 1) to run FineStructure algorithm to build a tree (as recommended for large dataset, we

performed 10,000,000 burn-in and runtime MCMC iterations) (Lawson *et al.* 2012) and 2) to perform a PCA (PCA-CP) with R (R Core Team 2015).

ROH detection

To estimate inbreeding we defined the ratio F_{ROH} as the total length of an individual's genome covered by runs of homozygosity divided by the total length of the genome captured, here 2.7 X 10^9 base pairs. This estimation was found to be the most powerful method to detect inbreeding effects (Keller *et al.* 2011). We defined ROHs, based on recommendations from Howrigan *et al.* (2011), as long stretches with a minimum of 50 consecutive SNPs, allowing no heterozygous call and a maximum of 2 missing calls and found in a LD-pruned dataset. Keeping only individuals identified as FCs and who are not close relatives, we used all variants with a MAF over 5% (1,242,267 SNPs) and we removed SNPs with a multiple $r^2>0.50$ with any other SNP in a 50-SNP window (shifting every 5 SNPs), leading to a set of 205,335 SNPs. LD-pruning and ROH detection were performed with PLINK (Purcell *et al.* 2007).

Inbreeding depression analysis

We tested for an association of F_{ROH} with our traits using linear regression models. Sex, age, age^2 and an age by sex interaction were also fitted as covariates in the models. We also performed analyses with additional adjustment for population stratification by using ancestry-informative PCs obtained from the chunk count matrix. Controlling for a large number of ancestry-informative PCs entails the risk of removing true inbreeding effects on the traits, while not controlling for any is equivalent to denying that different ancestral groups might have different levels of inbreeding and different distribution of traits not necessarily caused by the inbreeding itself. Models with adjustment for 0, 3 and 10 ancestry-informative PCs were analysed. We set a p-value threshold for significance at 0.05. We are conscious that this choice is not very stringent considering all traits under study.

Results

Capturing individuals of French Canadian ancestry

PCA on pruned genotypes and on chunk count matrix gave similar results for the whole sample, highlighting differences coming from the most distant groups, i.e. individuals of Asian, African and European descent (Supplementary Figure 4.3, p.129). Focussing on individuals that declared a European ancestry (n=887), which was confirmed by the first PCA analysis, both PCA-CP and PCAgeno revealed the presence of a sibling pair (clear cluster observable on the second principal component (PC) for PCAgeno and on the third PC for PCA-CP (results not-shown)). Once one sibling was removed, PCA-CP shows genetic differences among the different groups of Europeans with East-European (people from Bulgaria, Germany, Hungary, Poland, Romania, Russia and Ukraine) and Italians clustering apart and also French individuals clearly differentiated from the FC individuals (Figure 4.1, p.119). PCAgeno allows for much less clear differentiation between the groups of different ancestral origin. With PCAgeno, the individuals are basically spread over 2 gradients, one led by a group of Italian individuals and the other formed by FC individuals (Supplementary Figure 4.4, p.130). Figure 4.1 also shows some scattering among FC individuals.

The use of the FineStructure algorithm to cluster populations confirms our conclusions on ancestral origins and goes one step further by establishing ancestrally related groups. As for PCA the most differentiated groups are the large continental group (Figure 4.2, p.120). Among individuals with a European descent, almost all Canadians are clustered together. By considering the place where participants were recruited (either Quebec City, Montreal and Saguenay-Lac-St-Jean), we further identify three FC groups fitting fairly well to sampling places. Moreover how those three groups are ancestrally linked on the tree is in line with demographic history. Saguenay underwent a regional founder event, thus making this regional population more differentiated and resulting in an earlier split on the tree. Meanwhile Montreal and Quebec have experienced a greater mixing and seem to have stayed more close to some Europeans (in particular French, information not shown). The tree specifically allows

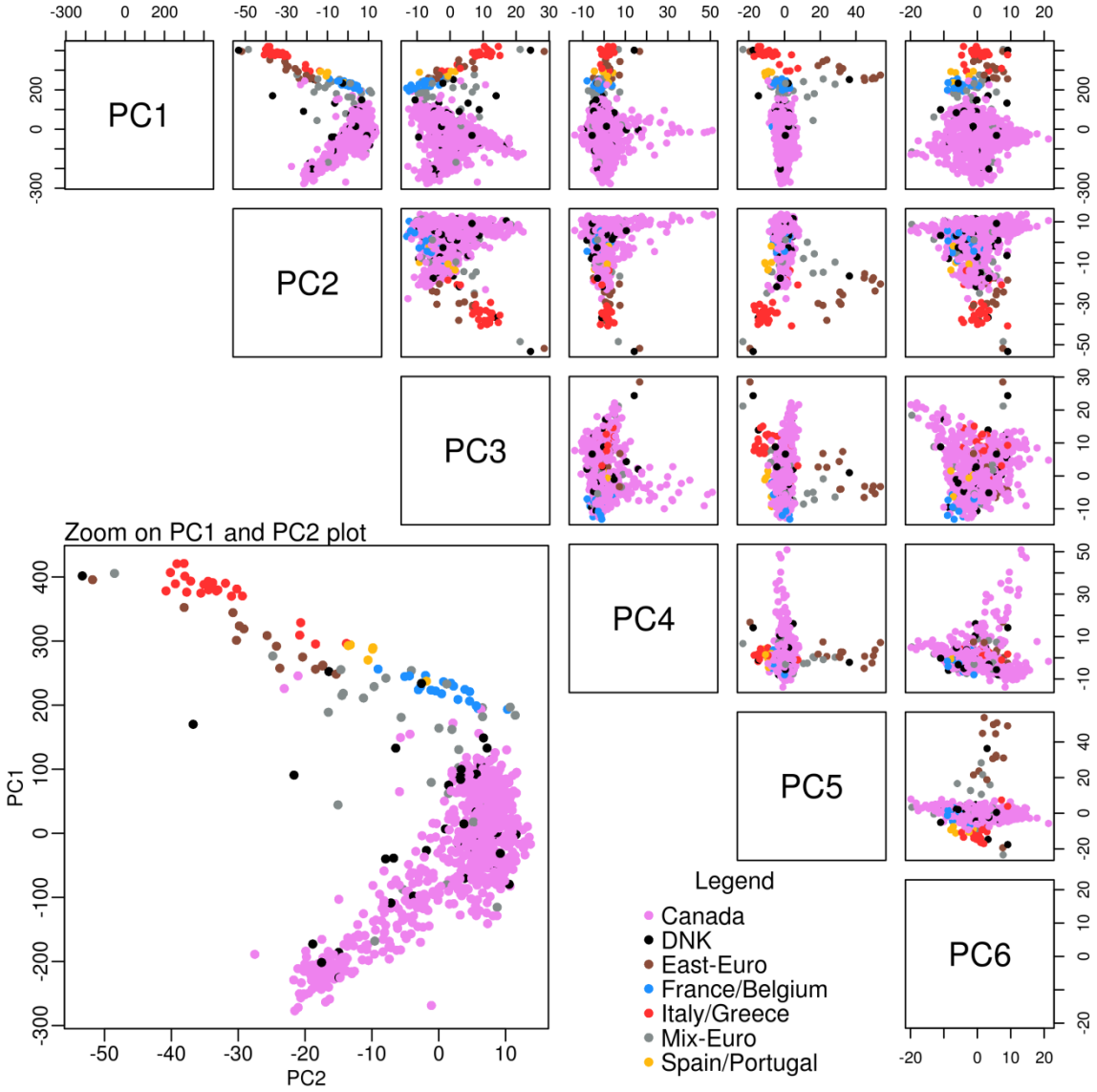

Figure 4.1 PCA on individuals of European descent

Plots of the first six principal components from the PCA performed on the chunk count matrix generated with ChromoPainter. Each point represents one individual and is colored according to the country of birth of the individual's grandparents. *DNK* Do not know, *East-Euro* Includes Bulgaria, Germany, Hungary, Poland, Romania, Russia and Ukraine, and *Mix-Euro* Admixed individuals from different European populations.

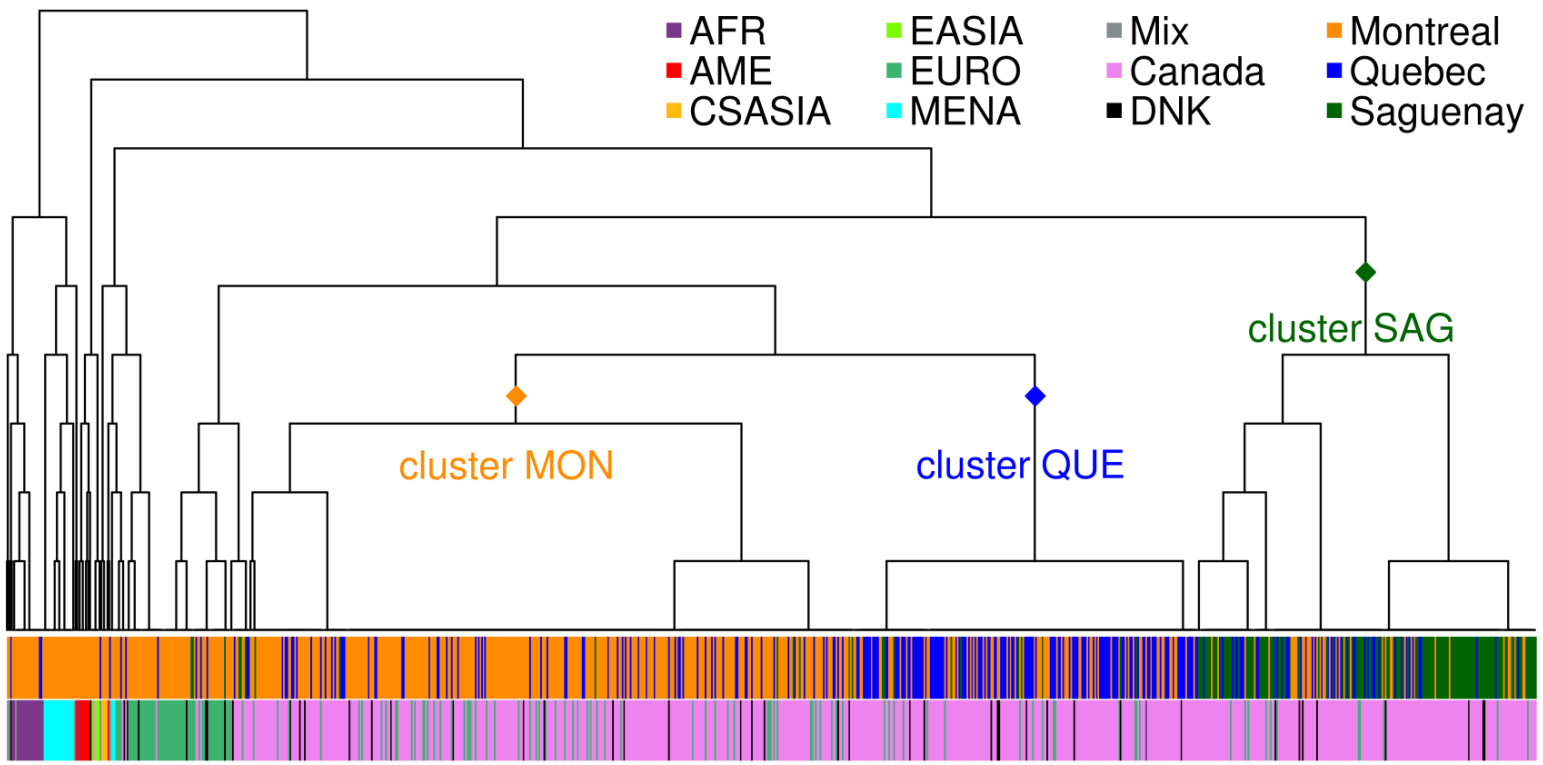

Figure 4.2 Population tree

Ancestral relationships between all individuals. Each bar represents one individual. First line under the tree is colored according to each individual's sampling place and second line is colored according to the grandparents' country of birth declared by each participant. The three FC branches (*MON* Montreal, *QUE* Quebec, *SAG* Saguenay) are shown. *AFR* Africa, *AME* America, *CSASIA* Central South Asia, *DNK* Do not know, *EASIA* East Asia, *EURO* Europe, *MENA* Middle East and North Africa, *Mix* Admixed individuals from different continental populations.

us to identify FC individuals that probably moved out from the regional population they ancestrally belonged to. We re-labelled FC individuals according to their identified ancestral group for subsequent analyses, such as how ROHs are distributed. PCAgeno restricted to FC only (n=726) also shows the presence of those three groups (see Supplementary Figure 4.5, p.131).

Homozygosity

Using Plink we identified ROHs for samples in the 3 FC populations. The distributions of total length of ROHs were significantly different across FC populations (Kolmogorov-Smirnov test p-values <0.05). We noted that levels of inbreeding tended to be higher on average in the Saguenay population, which has a smaller effective population size due to its demographic history of successive bottlenecks (Table 4.1). On average among all FC individuals the proportion of the genome located in ROHs was 0.0066, which is close to the fraction of the genome expected to be homozygous for a child of third cousins (F_{ROH}=0.0039). Only 2 individuals have a homozygous proportion of their genome such that their parents could have been first cousin or closer relatives (F_{ROH}>0.0625), while 47 individuals have a proportion higher than what is expected for offspring from second cousins (F_{ROH}>0.0156).

Table 4.1 Descriptive statistics for number of ROHs and proportion of genome covered by ROHs (F_{ROH})

Inbreeding measure		Minimum	Maximum	Mean	Median	SD
Number of ROHs	Montreal	2	52	9.15	9	4.40
	Quebec	2	18	9.17	9	3.02
	Saguenay	3	26	11.33	11	4.63
F_{ROH}	Montreal	0.0005	0.2058	0.0057	0.0036	0.0129
	Quebec	0.0007	0.0159	0.0049	0.0042	0.0029
	Saguenay	0.0008	0.0424	0.0096	0.0069	0.0083

Sample size: Montreal 325, Quebec 200 and Saguenay 201. *SD*: standard deviation

Association between inbreeding and traits

Supplementary Table 4.3 (p.132) presents the distributions of the traits studied. Overall the means in our sample of French Canadian are very close to the means in the CARTaGENE cohort. We could have found some discrepancies since our sample is much smaller and individuals were selected according to some cardio-metabolic features. We also observed large correlation coefficients (r>0.5) between a few traits, such as white blood cells and neutrophils, systolic and diastolic blood pressures, and waist to hip ratio and high density lipoprotein cholesterol (see Supplementary Table 4.4, p.134).

We tested for an association between F_{ROH} and the phenotypic traits with linear regression models adjusting for age, sex, age^2 and an interaction between age and sex. As shown in Table 4.2 (p.123), we found significant associations for height, body mass index, white blood cells (WBC) and neutrophils counts when we were not accounting for ancestry. The associations with height and BMI were still significant after adjustment for 3 or 10 ancestry-informative PCs. Associations of inbreeding with white blood cells and neutrophils counts were not significant anymore when ancestry was accounted for but associations were still suggestive and the same trend for effects was observed. Note that all significant associations, as well as many other non significant associations, were in the direction of a loss of fitness, i.e. smaller height, bigger BMI, lower WBC and neutrophils counts.

Although we could think that the effect on height is driving the effect on BMI, they are both weakly correlated (r=-0.063) and in fact BMI is highly correlated to weight (r=0.849) for which no significant effect was found. In Joshi *et al.* (2015), in addition to height they also found a significant association for forced expiratory volume in 1 second (FEV1). They explained that height was probably driving the significant association found for lung capacity since this last one was likely related to the trunk length, itself depending on the body height. However in our study lung function was not associated to inbreeding. For associations of inbreeding with WBC and neutrophils counts correlation is however a good explanation since both traits are highly correlated (r=0.918).

Table 4.2 Analysis of the association of the proportion of genome covered by ROHs (F_{ROH}) and various phenotypes

Trait	N	F_{ROH}		F_{ROH} corrected for 3 PCs		F_{ROH} corrected for 10 PCs	
		Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Anthropometric							
Height	721	-56.243	0.016*	-57.574	0.018*	-59.941	0.015*
Weight	695	28.380	0.605	35.328	0.540	30.832	0.596
BMI	695	39.106	0.043*	42.600	0.036*	40.952	0.046*
Waist to hip ratio	698	0.241	0.401	0.211	0.484	0.262	0.384
Peripheral blood pressure							
Systolic	668	-50.955	0.382	-82.049	0.179	-79.206	0.197
Diastolic	668	-26.342	0.505	-29.013	0.484	-29.350	0.482
Lung function							
FEV1	583	-2.437	0.394	-1.364	0.647	-1.768	0.554
FEV1/FVC	583	-1.592	0.484	-0.750	0.752	-0.998	0.676
Haematology							
Red blood cells	720	-0.568	0.667	-1.365	0.323	-1.385	0.318
White blood cells	720	-15.038	0.035*	-13.035	0.079	-13.810	0.064
Haemoglobin	720	-42.894	0.230	-43.133	0.250	-41.896	0.265
Haematocrit	720	-0.133	0.203	-0.142	0.195	-0.140	0.202
MCV	720	-17.269	0.262	-2.524	0.874	-1.561	0.923
MCH	720	-4.781	0.420	1.252	0.839	1.718	0.782
MCHC	720	4.154	0.855	12.107	0.611	13.292	0.579
RCDW	720	1.042	0.806	-0.564	0.899	-0.784	0.861
Platelets	719	-34.222	0.867	-103.085	0.626	-130.677	0.539
Lymphocytes	717	-1.930	0.464	-1.337	0.628	-1.424	0.608
Monocytes	717	-0.642	0.271	-1.006	0.099	-1.005	0.101
Neutrophils	717	-11.551	0.036*	-10.087	0.079	-10.818	0.061
Eosinophils	717	-0.307	0.497	-0.217	0.647	-0.263	0.582
Basophils	717	-0.137	0.371	-0.044	0.784	-0.071	0.658
Biochemistry							
Triglycerides	726	-4.193	0.346	-3.416	0.463	-3.044	0.516
Total cholesterol	726	-1.452	0.703	-2.011	0.613	-2.485	0.533

HDL cholesterol	726	0.032	0.982	-0.350	0.816	-0.378	0.802
LDL cholesterol	701	0.864	0.801	0.178	0.961	-0.299	0.934
HBA1c	722	-0.001	0.952	0.002	0.924	0.002	0.924

Analyses were adjusted for sex and age and performed either without adjustment for ancestry or with adjustment for ancestry by including the first 3 or the first 10 ancestry-informative principal components (PCs). *BMI*: Body mass index, *FEV1*: Forced expiratory volume in one second, *FVC*: Forced vital capacity, *HBA1c*: Glycated haemoglobin, *HDL*: High density lipoprotein, *LDL*: Low density lipoprotein, *MCH*: Mean corpuscular haemoglobin, *MCHC*: Mean corpuscular haemoglobin concentration, *MCV*: Mean corpuscular volume, *RCDW*: Red cell distribution width. .* p-value < 0.05

Discussion

By identifying FC individuals among a group of diverse individuals, our study confirmed that FC has an important population structure. Regional populations of the province of Quebec had already been studied and described (Roy-Gagnon *et al.* 2011). Here we highlighted the differentiation of FC individuals from their French source population, which had not been shown as clearly in previous studies. Also considering all SNPs and the whole haplotypic structure was the key to further observing a differentiation between the two metropolitan regions (Montreal and Quebec City). It has been shown that studying haplotypes instead of genotypes frequencies yield greater power to discriminate ancestrally related groups (Lawson and Falush 2012), and this greater power enabled us to clearly identify the participants' FC ancestry. The ultimate thing that could have helped us to confirm the identification of FC individuals and refine their ancestral origin is the use of genealogical information. This information is available through the BALSAC population repository for a part of CARTaGENE participants and thus, potentially for individuals in our sample (Awadalla *et al.* 2013; BALSAC 2014a).

In a previous study, we have shown that FC individuals tend to share an important number of common ancestors, an information that was derived from extensive genealogical data (Gauvin *et al.* 2014). Here, we found varying patterns of inbreeding in three FC regional populations and these patterns matched the demographic history of FC as well as the accumulation of common ancestors. A previous study in the FC population clearly showed that those patterns of inbreeding are highly correlated to expectations from genealogical records for FC (Roy-Gagnon *et al.* 2011).

Having assessed the extent of inbreeding in our group of FC, we investigated its impact on a range of physical, haematological and biochemical traits. The most significant association found was with height. The fact that more inbred people tend to be smaller, in our study it is a loss equivalent to about 3 cm for offspring of first degree cousins compared to slightly inbred offspring of third cousins or more distant relatives, is well documented and has been the subject of a large meta-analysis (McQuillan *et al.* 2012). As mentioned in the introduction,

height is a highly heritable trait and this is probably an important factor in explaining inbreeding depression over that physical characteristic. The effect on BMI was documented to our knowledge only once in a children cohort (Fareed and Afzal 2014) and as we explained could not be driven by height since both are poorly correlated.

We consider the standard threshold of 0.05 for significance although we are conscious that this may be a too liberal approach. If we consider a Bonferroni adjusted p-value, none of the associations would be regarded as significant. A larger sample size would be needed to untangle which associations are more likely to be real and which are not. However in the light of previous studies on inbreeding depression, we still think our study is suggestive of effects on anthropomorphic and potentially also on haematological traits. The adjustment for ancestry is important since some of the effect on the traits may be driven by ancestry differences. However, inbreeding is obviously a component of allele and haplotype frequencies differences (i.e. ancestry).

The new association found for WBC and neutrophils with inbreeding was not, to our knowledge, ever reported. However a study has shown that consanguinity is a factor in the susceptibility to infectious diseases in humans (Lyons *et al.* 2009) and studies on animals showed a correlation between consanguinity and pathogenic infections (Ilmonen *et al.* 2008; Murray *et al.* 2013; Hoffman *et al.* 2014). Since WBC are the immune system cells, and therefore the ones fighting off infections, the decrease amount of WBC could explain more directly the increased susceptibility to infections. There is also a study which showed that genes differentially expressed between inbred lines of drosophila and noninbred lines were enriched for genes involved in metabolism, like defense mechanisms (Kristensen *et al.* 2005). It is also worth noting that WBC and neutrophils are highly correlated, as reported in the results, since neutrophils make up to 60% of WBC. This also explains why inbreeding is associated with both traits. These two associations' results were not significant anymore when adjusted for ancestry but the effects were still in the same direction.

Our study did not replicate previously reported associations between inbreeding and cardio-metabolic traits. However, for the majority of the traits studied, the effects found were in the

same direction as the previously reported associations. The only exception was mean corpuscular haemoglobin corrected for ancestry. Since heritability is a population parameter (Visscher *et al.* 2008), inbreeding depression effects are probably population-specific, which could explain the differences found. Moreover heritability for height might be more consistent across different populations, a reason why we replicated that result, while it might not be the case for the other traits under study. Inbreeding and selection both have an influence on genetic variance in a population and therefore they also influence heritability (Visscher *et al.* 2008).

As reported in a very recent study (Joshi *et al.* 2015), it may also be that previous studies did not have enough power and some of their results may have been false positives. In this study they actually found a signification association between inbreeding and height, education attained, cognitive ability and forced expiratory volume in one second (FEV1) in a meta-analysis of multiple cohorts leading to a sample of over 350 000 individuals. In our study we did not have access to information on cognitive ability and associations with FEV1 and education level (not presented) were not significant. Another study performed on a large Finish cohort, about 5500 individuals, did not either replicate the results about blood pressure but reported significant associations with height and education.

Another explanation for this non-replication of certain results relies on the mechanisms leading to the inbreeding identified in the FC population. If we suppose that inbreeding created over one or two generations has a stronger and more direct impact than the equivalent inbreeding created over a large number of generations, this would mean that FC individuals, having mostly inbreeding stretches derived from the accumulation of distant common ancestors, would be less influenced by inbreeding depression. Whether inbreeding enables the activation of deleterious recessive alleles, which have been shown to be more prevalent on long ROH (Szpiech *et al.* 2013), is still a relevant question. As Ruderfer *et al.* (2015) said, population genetic and ancestry characteristics could also be a determinant in the role of recessive variants and homozygosity in the variation for different traits.

An important strength of our study is that inbreeding was assessed by the mean of genetic information. As stated in Bittles and Black (2010) a clear assessment of consanguinity must be done before any aspect of health can be associated to a purported presence of inbreeding. Indeed for many studies on the subject the estimation of inbreeding or homozygosity was done using either information reported by participants, genealogical data or using sparse genetic markers. Information provided by participants may not be as reliable as we think knowing among others, that inbreeding can raise social or cultural issues and it does not reflect the full extent of inbreeding if it is limited to a status of consanguineous or not. Genealogical data may provide more accurate and in depth information, however it can be error-prone, it is in most populations fairly limited (2-3 generations deep) and more importantly it does not account for stochastic variation in the inheritance process. Thus, performing a genetic evaluation of the actual realized inbreeding is getting very important. Yet many studies did that genetic inference but with microsatellites markers, which may provide an incomplete or rough portrait of inbreeding. On our side, we had very dense genome-wide SNPs data, which provide trustworthy results.

Using a very dense panel of SNPs for a group of FC deeply phenotyped, we investigated how inbreeding depression influenced quantitative traits. Consistent with previous published work, we found that height was associated with inbreeding. We also brought to light a potential new association with WBC however this should be interpreted with caution because we are the first to report the association and its statistical significance falls when the association is adjusted for ancestry. Our study is one of the rare studies to investigate associations between genetically assessed inbreeding depression and a large number of health-related traits, using a considerable size sample. On the basis of our results, we believe that further investigation of inbreeding depression is warranted and needed to obtain greater evidence that inbreeding depression has an impact on human health in different populations so that different demographic parameters and environmental factors are considered as well. Additional work in this field will help to elucidate the respective role of accumulation of rare recessive alleles versus common ones, to understand the impact of assortative mating and to better understand the evolutionary forces shaping our genomes.

Supplementary Figures and Tables

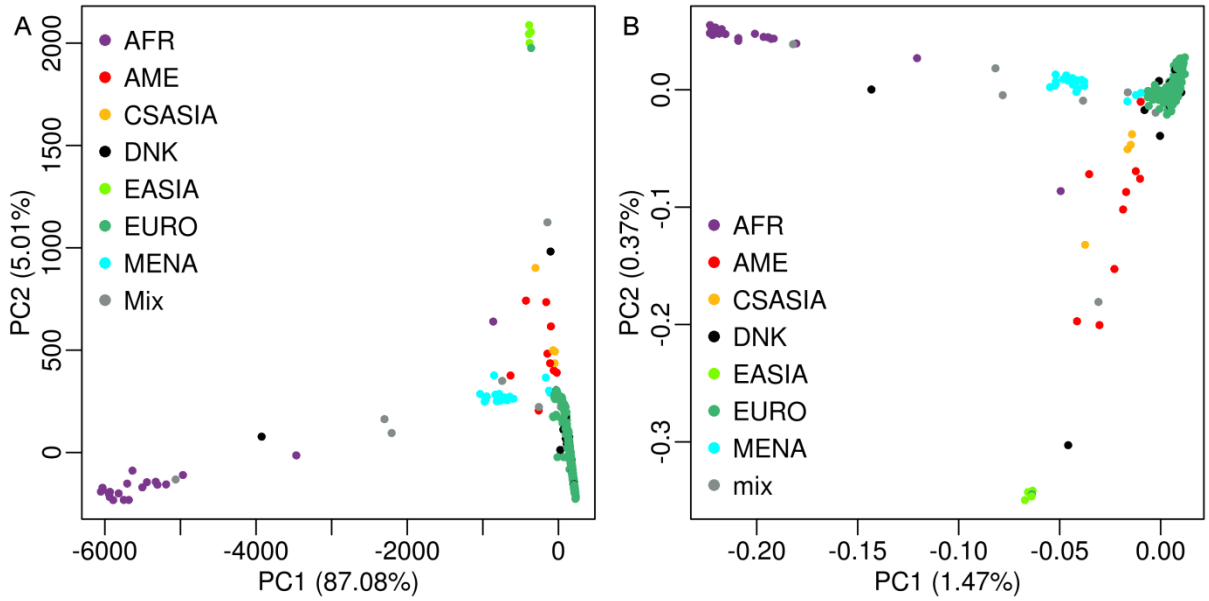

Figure 4.3 PCAs on all individuals

Plots of the first two principal components from the PCA performed on A) the chunk count matrix generated with ChromoPainter and B) LD pruned genotypes. All individuals are colored according to the grandparents' country of birth declared by each participant. *AFR* Africa, *AME* America, *CSASIA* Central South Asia, *DNK* Do not know, *EASIA* East Asia, *EURO* Europe, *MENA* Middle East and North Africa, *Mix* Admixed individuals from different continental populations.

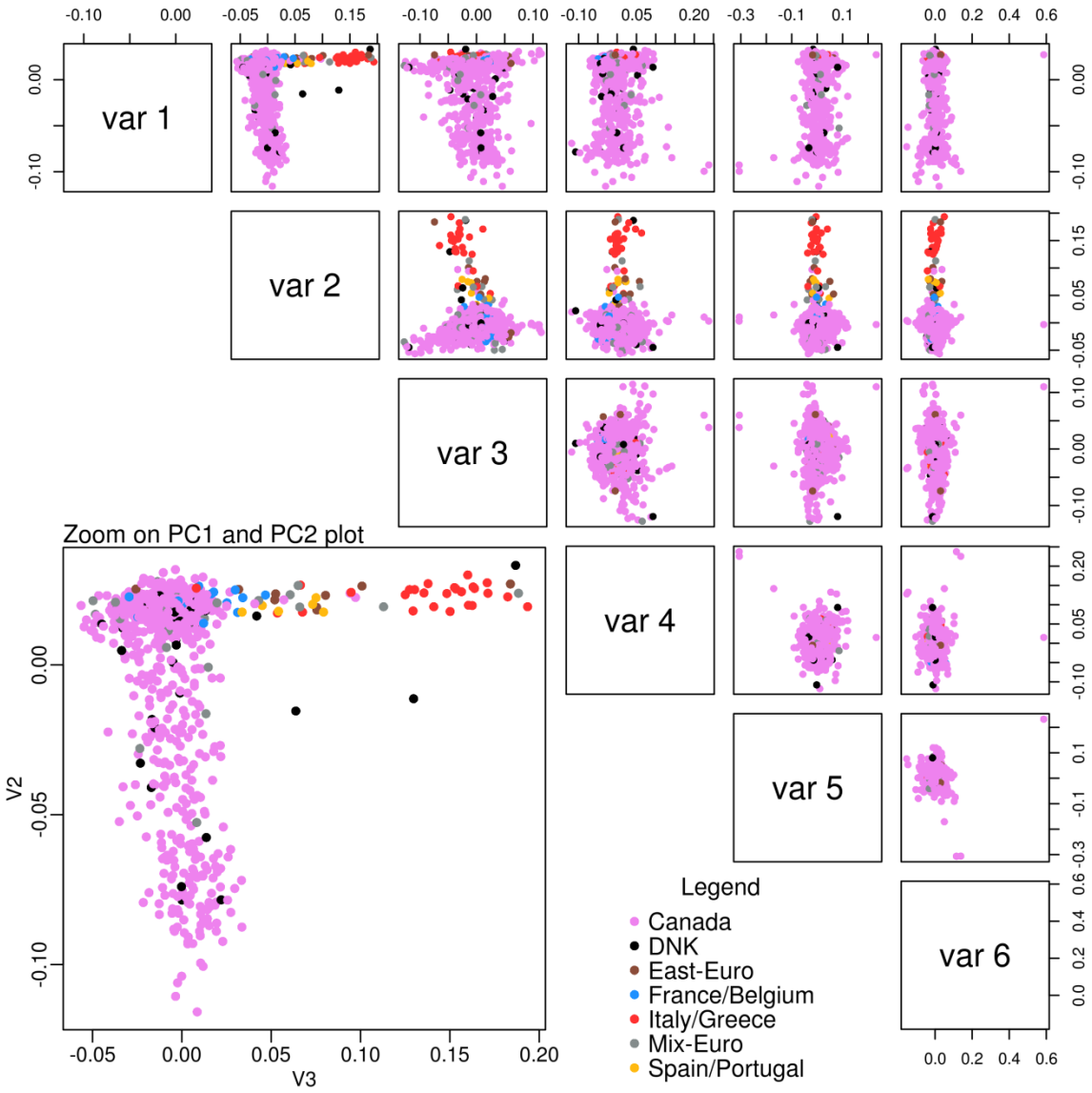

Figure 4.4 PCA on genotypes from individuals having a European descent

Plots of the first six principal components from the PCA performed on LD pruned genotypes. Each point represents one individual and is colored according to the country of birth of the individual's grandparents. *DNK* Do not know, *East-Euro* Includes Bulgaria, Germany, Hungary, Poland, Romania, Russia and Ukraine, *Mix-Euro* Admixed individuals from different European populations.

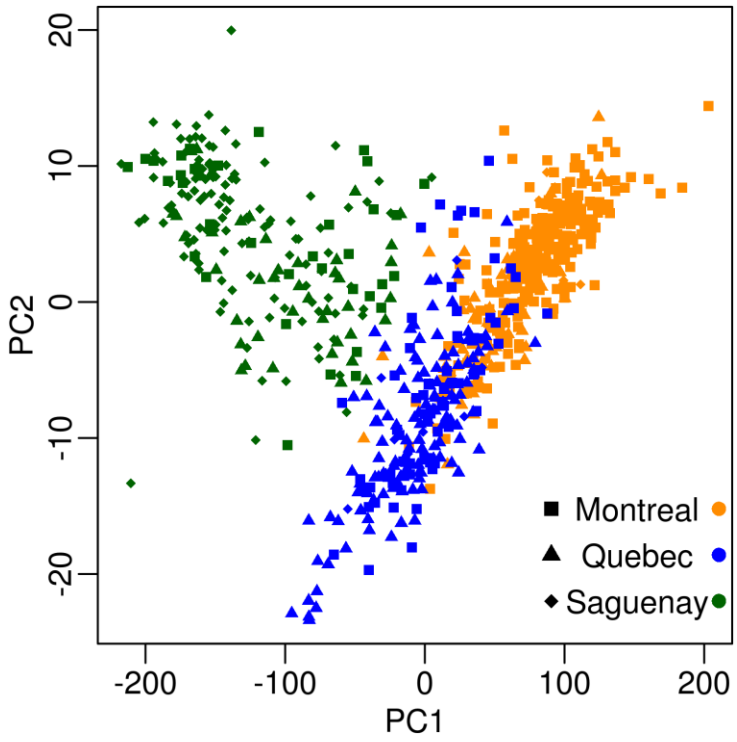

Figure 4.5 PCA on genotypes from French Canadians individuals

Plots of the first two principal components from the PCA performed on the chunk count matrix generated with ChromoPainter only for FC individuals (n=726). Individuals are colored according to their identified ancestral group and dot shapes are showing where they were sampled.

Table 4.3 Summary statistics for all phenotypes in French Canadians

Traits (units)	N	Minimum	Maximum	Mean	SD
Anthropometric					
Height (cm)	721	143	193.3	166.8	9.2
Weight (kg)	695	42	142.6	74.9	16.1
BMI (kg/m²)	695	15.1	52.7	26.8	5.2
Waist to hip ratio	698	0.5	1.2	0.9	0.1
Peripheral blood pressure					
Systolic (mmHg)	668	85	206	124.9	18.1
Diastolic (mmHg)	668	46	125	72.7	11.1
Lung function					
FEV1 (L)	583	0.8	5.87	3.0	0.8
FEV1/FVC (%)	583	36.3	99.6	78.8	8.0
Haematology					
Red blood cells (10^{12} cells/L)	720	2.91	5.71	4.5	0.4
White blood cells (10^9 cells/L)	720	2.96	16.3	7.0	2.0
Haemoglobin (g/dL)	720	100	175	138.5	12.2
Haematocrit	720	0.295	0.519	0.4	0.0
MCV (fl)	720	77.1	124.2	91.0	4.2
MCH (pg)	720	25.9	42.8	31.1	1.6
MCHC (g/L)	720	320	360	341.0	6.1
RCDW (%)	720	10.4	18.9	13.4	1.1
Platelets (10^9 cells/L)	719	27	494	246.1	56.3
Lymphocytes (10^9 cells/L)	717	0.66	11.7	2.0	0.7
Monocytes (10^9 cells/L)	717	0.15	1.2	0.5	0.2
Neutrophils (10^9 cells/L)	717	1.5	12.2	4.3	1.5
Eosinophils (10^9 cells/L)	717	0	1.1	0.2	0.1
Basophils (10^9 cells/L)	717	0	0.23	0.0	0.0
Biochemistry					
Triglycerides (mmol/L)	726	0.3	10.3	1.8	1.3
Total cholesterol (mmol/L)	726	2.5	9.4	5.1	1.1
HDL cholesterol (mmol/L)	726	0.41	3.21	1.3	0.4
LDL cholesterol (mmol/L)	701	0.7	7.4	3.0	1.0
HBA1c	722	0.046	0.13	0.1	0.0

BMI: Body mass index, *FEV1*: Forced expiratory volume in one second, *FVC*: Forced vital capacity, *HBA1c*: Glycated haemoglobin, *HDL*: High density

lipoprotein, *LDL*: Low density lipoprotein, *MCH*: Mean corpuscular haemoglobin, *MCHC*: Mean corpuscular haemoglobin concentration, *MCV*: Mean corpuscular volume, *RCDW*: Red cell distribution width.

Table 4.4 Highly correlated traits

Traits		Correlations
Haemoglobin	Haematocrit	0.9787
MVC	MCH	0.9326
Total cholesterol	LDL cholesterol	0.9191
White blood cells	Neutrophils	0.9178
Red blood cells	Haematocrit	0.8825
Red blood cells	Haemoglobin	0.8526
Weight	BMI	0.8490
Systolic BP	Diastolic BP	0.7600
Height	FEV1	0.6820
White blood cells	Lymphocytes	0.5876
White blood cells	Monocytes	0.5513
Waist to hip ratio	HDL cholesterol	-0.5574

All traits with $|r| > 0.5$. *BP*: Blood pressure

Chapter 5:

Discussion

5.1. Summary and discussion of main findings

Isolated populations are very useful for genetic research and the French Canadian population can help us to decipher the biological mechanisms behind diseases and phenotypical differences. In this thesis, we studied the patterns of genetic relatedness across and within individuals resulting from various genealogical relationships. In a first project, we inferred patterns of IBD sharing in a sample of French Canadian individuals coming from seven ethno-cultural sub-populations for which we had access to genealogical records. In a second project, we presented an efficient tool to analyze extended genealogical information. In the last project, we investigated a larger sample of individuals having a French Canadian descent to study the impact of inbreeding depression on various health phenotypes.

5.1.1. Family relationships and genetics

The genetic material is being reshuffled over generations every time new offspring are conceived and DNA strands keep track of all parental unions that led to these new remixes. With the advent of genotyping technologies, we were able to better characterize genetic variations and different methods have emerged in order to use the genetic information to infer chromosomal segment that are shared IBD between a pair of individuals. A common way to evaluate the power, the false-positive rate and the accuracy for a method detecting IBD segments rests on performing a simulation study, which preserves the true information to be detected.

In our first study, we proposed an evaluation framework in which we had genetic data and the genealogical information relating to it. We were able to assess the global accuracy of these methods by using a panel of French Canadian individuals. We performed the IBD detection with five methods and using also two phasing techniques as phasing is likely to impact the subsequent inferences for methods using phased data. We found that indeed phasing could greatly influence the ability to accurately infer IBD segments and that one of the best methods was capitalizing on a sampling of multiple phasing results. We obtained high correlations

between total length of IBD segments and relatedness measured using genealogical data (see Table 2.1 p.59). Using the IBD information generated we investigated how it was related to different genealogical characteristics. We observed that pairs of individuals having an inbred common ancestor were having significantly more IBD sharing than pairs who did not have that kind of common ancestor (see Figure 2.7 p.71). Total length of IBD sharing was also reflecting other characteristics of relatedness, such as distances to common ancestors. We found IBD sharing explained 39% of the variance in the distance to the nearest LCA (see Figure 2.2 p.61). At the population level, since our samples were coming from seven sub-populations of the French Canadian population, we were able to study the influence of various demographic features. Patterns of IBD sharing as well as average sharing were different across the sub-populations, consistent with the French Canadian settlement history. Overall IBD patterns observed were, as expected, in line with the complex relatedness present among French Canadians and they could probably be informative about the types of distant relationships in a sample from a founder population like Quebec.

During our study of the Quebec population, we recognized the need for an effective software for genealogical analysis with all the basic descriptive functions as well as functions of common use, such as for gene-dropping simulations. Therefore our second project combines the presentation of an R package for the analysis of genealogical data and a simulation study to investigate IBD sharing of alleles and chromosomal segments. R is a popular software environment used for statistical computing used in a wide range of scientific fields including genetic research (Foulkes 2009). Thus, it was an obvious choice to extend and implement GENLIB as a package within the R environment. We used the same panel of individuals as in the first study. We made a detailed description of the genealogical dataset using functions from our package. The whole genealogical corpus, coming from BALSAC population database, includes over 40 000 individuals, spans 18 generations and is provided with the package for the benefit of its users. We described how complete, complex and different the genealogies from the seven populations are, although they all have common distant ancestors.

We ran simulations to compare odds of IBD sharing for similar kinship levels that have arose from very different genealogical links. In a first scenario we looked only at the probability of

sharing one allele IBD. The chances of sharing an allele IBD depend on the distance separating the individuals through a common ancestor and the type of ancestor, which is defined according to the linking paths. Briefly, the closer a common ancestor is the higher the odds of IBD sharing are and these odds increase with the number of different paths. In a second scenario, we studied IBD sharing for chromosomal segments of varying lengths. We looked at the probability that a segment was transmitted intact through a common ancestor to two of its descendants. Recombination comes into play with other factors and influenced the odds of sharing a segment IBD. By comparing two pairs of individuals from two different sub-populations we highlighted how the distance and number of common ancestors influenced the IBD probabilities and how these features can help unravel which ancestors are more or less likely to have transmitted segments. With its implementation within the R environment and the new functions we implemented, the GENLIB package is an ideal tool to analyze information coming from genealogies and to combine it with genetic information.

Knowing that French Canadians have so many common ancestors and different levels of relatedness, it was clear they were also prone to be inbred although it was probably happening in a distant fashion. In our third study, we investigated how inbreeding depression could impact health traits for individuals having distant consanguinity. Inbreeding depression is a well documented phenomenon that influences various traits making them less favourable in most cases. In order to achieve our investigation we used a sample of participants from CARTaGENE Project. Since this biobank is population-based, we first had to identify all individuals having a French Canadian origin from those whom did not or had only a partial one. We used the latest tool to study the ancestral origin of all the individuals. We observed, as expected, that French Canadians are genetically close to Europeans and even closer to French while being still quite differentiated. Overall, the analysis of the ancestral origin of CARTaGENE participants was consistent with the information provided by participants about their origin. We focused on individuals identified as persons having a French Canadian ancestry and we estimated their level of inbreeding by looking at the ROHs. Homozygosity rates across French Canadian varied according to their region of origin. We investigated the correlation between homozygosity and different phenotypes. We found evidence of inbreeding

depression effects on height and BMI and potential new associations with white blood cells and neutrophils counts.

Our contributions to the field are numerous. Our first study added to the knowledge of population structure within the French Canadian population and to the description of the great variety of types of relatedness present. The second study revealed how these relationships' characteristics can impact the transmission of genetic material in descendants. We also presented in detail the functionalities of an R package, called GENLIB, to analyze extensive genealogical data. This tool allows an efficient and easy integration of different types of data whilst being built in an environment facilitating new developments, such as the addition of new functions. The GENLIB package is a highly valuable tool to study extended genealogies. Our final study added evidence to consider inbreeding depression as having an effect on height and BMI; also we found a potential novel association with white blood cells count and in particular neutrophils. This study also demonstrated how we can easily identify isolated populations, such as the French Canadian population, among a group of individuals with various ancestral backgrounds.

5.1.2. Strengths and limitations of the studies

The sample used for the two first studies is a panel of French Canadians coming from different regional or ethno-cultural populations. Although the sample allows investigating the population structure within the French Canadian population, as it includes seven sub-populations, the panel does not cover the whole province and sample sizes of each subpopulation are not quite large ranging from 16 to 22 individuals. Sampling was performed by paying particular attention to the geographical origin of participants, which is expected to ensure a good representation. Participants were selected preferentially if they had at least one parent born in the region before 1960 or were themselves born in the region before 1960. To ensure that participants' ancestral origin was specific to a region, we could also have put a more stringent criterion, like birth or marriage within the region for 2 or 3 generations preceding them. Increasing the sample sizes could also have increased the evidence that

samples were representative of their own region. In this regard the Quebec Reference Sample now includes 2 more regions (*Abitibi* and *Outaouais*) and another ethno-cultural group from Gaspesia, as well as a few more individuals for previously sampled regions (QRS 2015).

In the first study we focused on the study of IBD patterns in the French Canadian population. For this study we had access to extensive complex genealogical data allowing tracing back French Canadians' ancestors up to the 17th century. The assessment of IBD inference methods is typically done in a framework using simulations. Here we evaluated how well IBD sharing inferred was correlated to the expected sharing computed from the genealogical data, which was a premiere, at least according to our knowledge.

Before making any comparisons, the detection of the IBD segments needs to be done and its quality depends, among other things, on the genotypic data accuracy. Genotype errors can significantly reduce the power to detect IBD segments; this is why three out of the five methods we used can account for these errors. However having a stringent data quality control remains important so we applied one to our dataset in order to remove low quality variants. The capacity to accurately infer IBD segments also depends on the amount of IBD sharing present among samples. In Table 2.1 (p.59), we observed that the correlation within populations between IBD sharing and kinship coefficients tended to be higher, as expected, for the populations having closer relationships. More specifically we noted that some methods were underestimating IBD sharing, i.e. almost none or zero IBD segments, for pairs of individuals from populations with low levels of kinship. Meanwhile for some populations (Saguenay, Quebec and Loyalists) most methods were systematically inferring more IBD sharing than expected from genealogical information for some pairs of individuals. This may suggest that some additional genealogical information would have provided a more accurate picture for some allegedly distantly related individuals.

As presented in detail in the introduction (see section 1.4.2 p. 35), phasing is also likely to affect the detection of IBD segments. This is the reason why we used two phasing algorithms to observe their influence and to use the most effective method. One thing we did not do is to compare the regions detected as IBD across methods. We used the total length of IBD

segments to compute the proportion of the genome located in IBD segments without looking at whether these regions were the same or not between the different methods. This could have been another criterion to compare methods and we probably would have seen some discrepancies especially for the endpoints of the shared haplotypes since they are tricky to define. Segments' breakpoints could also have been put in relation to published estimates of recombination rate along the genome to see how well they fit but it was beyond this project goal. Another approach that we could have taken, would have been to look at the distributions of segment sizes and how these lengths are compared to expectations based on relatedness. Finally, our evaluation could have been improved if we did explore the parameters optimization of the methods used, which we did not do.

Our study is an additional example putting forward the importance of considering relatedness in a sample before studying it. The high correlation that we observed between genealogical information and IBD sharing, over the wide range of remote relatedness present in our study population, further demonstrates the usefulness of genomic IBD detection to capture even complex relatedness involving inbreeding. Our findings can guide the interpretation of IBD sharing results in other populations not having access to genealogical data.

The second study used the same panel for different purposes. The goal of this second project was to present the GENLIB package, originally developed to work with the BALSAC database and which we implemented in a new environment easily accessible. Most functions were already programmed for a proprietary software, S-PLUS (S-PLUS 8.2 Copyright 2008, TIBCO Software Inc.), resulting in a limited use by people outside BALSAC project and by people, who did not have access to the software. This is one reason why we thought this package could benefit a lot from being translated into the R environment (R Core Team 2015). R is a free software environment for statistical computing, meaning that users can run it, study it, change it and redistribute it with or without modifications as long as it is free of charge. R also offers a wide range of complementary functions through all other packages developed. In this new implementation we added new functions and improved some others. In other words, the new GENLIB package, which we believe is an improvement over existing softwares for

genealogical analysis, is user-friendly and flexible thereby encouraging its use and the contribution of other users.

However the package could still be improved in two ways; first by optimizing the computation algorithms and second by adding other relevant functions. We did compare computation times for certain functions for GENLIB and other softwares and it was found that GENLIB is not as fast as others can be. Therefore to have the best performance on all aspects we should work on optimizing the calculation algorithms. We also noticed that in order to ensure that the package will become popular among people who use pedigree and genealogies; it would probably need functions to import formats of genealogical data other than the specified format. Additional functions to easily extract information from extensive genealogical data were suggested but are not implemented yet.

Apart from presenting the GENLIB package with a detailed description of the genealogical corpus included within the package, we performed a simulation study. We did it using only two pairs of individuals and this may seem too little however the goal was only to compare pairs with very similar kinships, which were resulting from different ancestral links. We chose pairs of individuals with kinship value slightly larger than third degree cousins but we could also have picked additional pairs of individuals less related and coming from other sub-populations than the ones selected. A more elaborate study could have revealed other population characteristics influencing IBD sharing patterns among a population. For example, we could have selected a chromosomal region known to be shared by more than two individuals coming from various sub-populations and investigate which ancestor it was coming from.

The last study was performed on another sample of French Canadians coming from the CARTaGENE biobank. The first step was actually to identify the French Canadians since CARTaGENE is a population-based project, which recruited Quebecers aged between 40 and 69 years old without any restriction on their ancestral background. To achieve this identification we used the latest method, ChromoPainter algorithm, which is an improvement over traditional methods such as PCA on genotypes (Lawson and Falush 2012). Generally,

groups identified fitted declared ancestry and knowledge about population structure. However, to make sure we kept exclusively French Canadian individuals we further restricted our selection to individuals with at least 3 grandparents born in Canada. We ended up with a sample size of 727 French Canadian individuals, which is great when we compared to most inbreeding depression studies but quite small in comparison to a very recent meta-analysis performed on multiple cohorts for a grand total of over 350 000 individuals. Thus we still had more power than most studies investigating inbreeding depression on several traits at once. Considering our sample size and power to identify association we may miss some associations and effect sizes may be inflated and less precise. We also chose not to perform any p-value adjustment, like Bonferroni correction, even though we tested 27 different traits. We are aware that having 4 associations significant at a 0.05 level is not far from what we could expect by chance. Therefore our results should be interpreted with caution and further investigation is needed for the novel associations found concerning white blood cells count. However, we did replicate the height association, an association which gained evidence through different studies on large cohorts (McQuillan *et al.* 2012; Verweij *et al.* 2014; Joshi *et al.* 2015) and we obtained similar results as in a large meta-analysis recently published (Joshi *et al.* 2015).

Failure to replicate some previous findings in our cohort does not necessarily invalidate the original results since effects can be population-based. One major strength of our work is that we had very detailed phenotypes. Another important point is that we were able to perform a good evaluation of inbreeding as we had very dense SNPs data. We decided to adjust our regressions for age and sex to get rid of extra variation in the phenotypes but we could have investigated the simple correlation between traits and inbreeding since it is unlikely that age and sex confound any associations. Also instead of looking to continuous traits only, we could have investigated inbreeding values among cases and non-cases of specific diseases. Concerning diseases, some participants have been diagnosed for certain conditions and take the relevant medications; however this information was not included in our analysis and could have biased our results. In animals inbreeding depression effects have been shown to change under different conditions (Kristensen *et al.* 2006; Ilmonen *et al.* 2008). Medication intake could have hidden associations or make them less strong.

5.2. Future perspectives

5.2.1. Identical-by-descent sharing

Since we performed our first study IBD detection methods continued to improve. Among others the authors of FastIBD released a new version (Browning and Browning 2013a) and new methods for IBD inference, which improve the accuracy, the precision and the resolution scale, were launched (Browning and Browning 2013b; Tataru *et al.* 2014; Rodriguez *et al.* 2015). The use of IBD patterns will probably change with the advent of next-generation sequencing (NGS). Apart from a probable better definition of shared segments boundaries, NGS will allow to investigate variants arising de novo on IBD segments. Using the numerous annotation tools that have emerged (Huang *et al.* 2008; Ng *et al.* 2009; Wang *et al.* 2010), we will be able to annotate these variants for function and assess their contribution to disease in risk haplotypes' carriers.

As most current detection methods consider only pairwise IBD, new approaches to investigate multiple haplotypes simultaneously have been developed (Gusev *et al.* 2011; He 2013; Qian *et al.* 2014). The attention is now focused on the use of these segments to uncover the genetic basis of multiple diseases (Vacic *et al.* 2014). However, IBD mapping may not be always the most powerful method (Westerlind *et al.* 2015) and new technologies may challenge its use. Nonetheless IBD mapping combined with other association analysis may also reinforce findings.

Although different genetic measures of relatedness can be used (Speed and Balding 2015), IBD has been proven useful to infer relationships (Huff *et al.* 2011). IBD sharing could also be used to develop relatedness measures incorporating the complementary information from both genomic and genealogical data. Recently, a new method for relationship detection based on IBD and good to identify, with NGS, relationships more distant than third cousins, was published (Li *et al.* 2014). The inference method, which uses NGS data, provides only a slight increase in power for relatedness inference as compared to genotype-based inference methods

since sequencing data come with additional technical challenges, such as an excess of pairwise IBD.

IBD sharing is also widely used to do demographic inference and to investigate patterns of admixture (Gravel *et al.* 2013). A study using IBD segments gave insights on the divergence between European and African populations and on the gene flow between both populations (Harris and Nielsen 2013). Increased interest in IBD sharing also means that the theoretical framework for the IBD process is further developed. As an example, Carmi *et al.* (2014) developed a model to derive the mean number of shared segments having a specific length and the distribution of the number of shared segments.

Following on our work it will be interesting to study the origin of CARTaGENE participants as they get genotyped or sequenced. We will be able to provide a more detailed picture of all the regional particularities. Ideally, if CARTaGENE Project goes on with a third recruitment phase it should be dedicated to more remote locations, such as *Iles-de-la-Madeleine*. A recent study on the British population uncovered very detailed population structure (Leslie *et al.* 2015). Although the Quebec population is much more recent there are definitely relevant demographic specificities to investigate within the province and in relation to its progenitor population.

5.2.2. Genealogical information

No matter their interest about it or their desire to know, all individuals are preserving traces of the demographic history prior to their existence; it is all about being able to unscramble it. In Quebec we are very fortunate to have a rich source of genealogical data and it is good news that BALSAC database will be merged with similar population databases, also concerning the Quebec population and covering different periods or types of records, under a unique infrastructure for historical microdata on the Quebec population (BALSAC 2014c).

Another type of information that will help furthering research efforts is electronic health records (EHR) (Jensen *et al.* 2012). After many decades of discussion about establishment of

EHR in Quebec, computerization of health information is still not a reality. Although some clinics locally propose EHR, all health information (disease outcomes, treatments, prescription drugs, laboratory and imaging) are not gathered in a unique record. Once this health structure integrating patients' data on all aspects will be created and functional, genealogical databases could potentially be linked to it. A recent systematic review highlighted the fact that genealogical databases are barely used in health care but when they are, benefits are multiple (Stefansdottir *et al.* 2013b). Known to contribute to a better understanding of familial and genetic factors in diseases, the use of computerized genealogies was shown to improve conventional screening methods, to result in a more comprehensive family history and to change cancer genetic counselling (Thorsson *et al.* 2003; Brewster *et al.* 2004; Cannon Albright 2006). Therefore in Quebec we could use genealogical information along with EHR in genetic counselling efforts, such as in Iceland (Stefansdottir *et al.* 2013a) or for research purposes to investigate familial clustering of diseases (Cannon-Albright *et al.* 2013).

Genealogies can also be used to look at epigenetic effects as some may be passed on through following generations (Jirtle and Skinner 2007). Epigenetic effects are changes in gene expression caused by external factors, such as the environment, which are not involving any change in the DNA sequence. While early life factors have been shown to have an influence on disease susceptibility (Gluckman *et al.* 2007), a recent study involving Swedish people reported a transgenerational response to stressful events, such as a nutritional deficiency (Bygren *et al.* 2014). Although epigenetic phenomenon is known since a while, transgenerational epigenetic inheritance is a phenomenon less documented and the marks it leaves along the genome still deserve a better characterization, particularly in humans (Daxinger and Whitelaw 2010; Grossniklaus *et al.* 2013).

To make a good usage of genealogical information, you need good tools. Therefore, there will be work to do to maintain the GENLIB package up-to-date. Although it was recently released, we already have a lengthy list of potential improvements that we could do. We already mentioned the function to import other genealogical data format to which we should add the development of new description functions and the inclusion of already known concepts such as the founders' uniform contribution number (Gagnon and Heyer 2001). There is also a great

amount of complementary information, such as event locations, dates, comments, which should be included somehow with genealogical data but the package is actually lacking of options to do it. Finally, functions to cut down the genealogies are presently pretty basic; therefore functions allowing the definition of groups based on complementary information would be an improvement.

We said in the above section that CARTaGENE would probably benefit from enlarging its sampling population. In this regard, since CARTaGENE data can be coupled with genealogical information from BALSAC, investigating the genealogies to get input on where the French Canadian individuals are ancestrally from could be an interesting way to know which regions deserve a better coverage, i.e. where participants in future recruitment phase could be sampled.

5.2.3. Public health and genetics

Genetic epidemiology brings its own methods to investigate biologic mechanisms of disease. The field ultimately contributes to provide an improved health and has a prominent place in the development of the "4P medicine". The 4P stands for Prevention, Prediction, Personalized treatments and Participation (Hood and Flores 2012).

For many of the people involved in medicine, the 4P approach is the future for care. With different screening programs and counselling, diseases can be predicted or prevented before problems arise. In Quebec, and in numerous countries, newborn screening, using tandem mass spectrometry, can detect various diseases such as tyrosinemia, a disease found among the French Canadian population (see Table 1.2, p.28). There are also carrier testing initiatives for couples considering having children and wanting to learn about their status. If they are both carriers for the same disease, genetic counselling is offered and options for pre-implantation or prenatal diagnosis are presented. The recent example of the pilot-project in the Saguenay region for carrier testing of four major recessive diseases is a great example. The population of the region has demonstrated an interest for the project, which is still under study, and the project has led to an increased demand for some genetic services outside the region (Pouliot

and Rousseau 2014). Originally these tests were offered in a cascade testing approach, which is targeting mainly relatives of diseased individuals, but they are now open to the whole population. Preliminary discussions to expand the project to all the Quebec population are underway to avoid creating inequity in the access to these services. The region of Montreal also had its specific genetic screening programs for Tay-Sachs and Beta-Thalassemia diseases, which have proved, on the long term, to highly impact disease occurrence (Mitchell *et al.* 1996). As new opportunities for screening at large are available, experiences like previous experiences may provide valuable insight and guidance.

Another aspect of the 4P medicine is the treatment personalization. The best known example of personalized treatment is the intake of warfarine (Marin-Leblanc *et al.* 2012). Warfarine dosage intake can be predicted with different polymorphisms of the *CYP2C9* and *VKORC1* genes to ensure that the patient has the best possible response to treatment. This last example is part of an emerging field, called pharmacogenetics, which gives importance to considering differential genetic susceptibility to drug effects and reactions. Genetic testing may help to define the appropriate treatment and may also improve drug development efforts since genetic factors could explain why some patients do respond while others do not. Staggering costs of drug development is an important public health issue; genetics could help to decrease them. Personalization also means the development of methods to deploy genomics based decision support (so that knowledgeable practitioner can use the information). As an example, a physician knowing a patient could be a poor metabolizer for a medication could improve his practice by considering alternative medications.

The last P of 4P medicine is participation; the participation of all citizens; patients, physicians, the medical community and the others. It ranges from education provided to them up to changes patients and citizens can do according to actionable information they obtained. Actionable information is information that is useful for improving the individuals' health, such as information about a genetic variant that can cause an adverse reaction when an individual is exposed to a drug. Citizens also constitute an important stakeholder in the development of new approaches to improve health (Knoppers and Joly 2007).

Another component of the development of a more pro-active medicine is better research resources, CARTaGENE being a good example. One of the initial study involving CARTaGENE has assessed the prevalence of chronic diseases in the cohort and found that the awareness (ratio of participants reporting a condition over participants affected by the condition) was pretty low as it ranged from 8% to 73% (Verhave *et al.* 2014). The research team also highlighted the factors influencing the awareness, among others being younger and affected by another condition, and furthermore investigated the proportion of participants reporting a condition but not meeting treatments targets (Verhave *et al.* 2014). The use of biobanks goes beyond research oriented towards biological aspects of diseases and their study can help to guide how and who should be targeted by specific medical care. Biobanks are also a public health tool. In addition to biobanks, EHR are increasingly playing a critical role in research and their use can be an important driver for EHR quality.

The integration of genetics as a part of general medicine is definitely, in my mind, one of the key step that could greatly improve medicine on many levels. However we have to remember that DNA is the profound core of individuality, meaning that personalized medicine has to be done with respect of choice, confidentiality and dignity.

As noted by Charles Scriver, more than twenty years ago, once the environment in under control and the behaviour is rather prudent the last thing that may cause diseases is the biology (Scriver 1988). The eminent paediatrician and geneticist concludes by saying that since genetics “are relevant to sick populations, because they constitute a way to anticipate sick individuals”, therefore “they are a form of public health”.

5.3. Conclusion

The French Canadian population has many interesting genetic features to investigate. In this thesis, we analysed genetic data to disentangle patterns of relatedness in some of its sub-populations. The genetic analyses were compared to expectations from extensive genealogical information. We found very good correlation between both genetic sharing and relatedness assessed with genealogical information. At the same time we highlighted characteristics of the identical-by-descent patterns among these French Canadian regional populations. Analysing genealogies require appropriate tools therefore we implemented an R package for the analysis of genealogical data. The tool was described in details by a complete description of a large genealogical corpus. A simulation study using the package functions, further revealed population characteristics influencing the genetic sharing. Finally, distant consanguinity rates were assessed in another sample of French Canadians and inbreeding depression phenomenon was investigated for numerous phenotypes. Significant effects of inbreeding were found for 4 phenotypes: height, body mass index, white blood cells and neutrophils count. We discussed well documented evidence for height while the association for white blood cells and neutrophils was documented for the first time. Nevertheless this last association could provide an explanation for an increased susceptibility to infectious diseases among people with a higher rate of genome-wide homozygosity.

References

Abifadel M, Varret M, Rabès J-P, Allard D, Ouguerram K, Devillers M, Cruaud C, Benjannet S, Wickham L, Erlich D, *et al.* 2003. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet* 34(2): 154–6.

Abney M, McPeek MS, and Ober C. 2000. Estimation of variance components of quantitative traits in inbred populations. *Am J Hum Genet* 66(2): 629–50.

Abney M, McPeek MS, and Ober C. 2001. Broad and Narrow Heritabilities of Quantitative Traits in a Founder Population. *Am J Hum Genet* 68(5): 1302–7.

Abney M, Ober C, and McPeek MS. 2002. Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *Am J Hum Genet* 70(4): 920–34.

Agarwala R, Biesecker LG, Hopkins KA, Francomano CA, and Schäffer AA. 1998. Software for Constructing and Verifying Pedigrees within Large Genealogies and an Application to the Old Order Amish of Lancaster County. *Genome Res* 8(3): 211–21.

Agarwala R, Biesecker LG, and Schäffer AA. 2003. Anabaptist genealogy database. *Am J Med Genet* 121C(1): 32–7.

Albrechtsen A, Moltke I, and Nielsen R. 2010a. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* 186(1): 295–308.

Albrechtsen A, Nielsen FC, and Nielsen R. 2010b. Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol* 27(11): 2534–47.

Albrechtsen A, Sand Korneliussen T, Moltke I, Overseem Hansen T van, Nielsen FC, and Nielsen R. 2009. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol* 33(3): 266–74.

Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, Gallacher J, Green J, Matthews P, Pell J, *et al.* 2012. UK Biobank: Current status and what it means for epidemiology. *Heal Policy Technol* 1(3): 123–6.

Arcos-Burgos M and Muenke M. 2002. Genetics of population isolates. *Clin Genet* 61(4): 233–47.

Avard D, Bucci LM, Burgess MM, Kaye J, and Heeney C. 2009. Public Health Genomics (PHG) and Public Participation : Points to Consider. 5(1): 1–21.

Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet J-P, Knoppers B, Hamet P, and Laberge C. 2013. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *Int J Epidemiol* 42(5): 1285–99.

Awan Z, Baass A, and Genest J. 2014. Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9): Lessons Learned from Patients with Hypercholesterolemia. *Clin Chem* 60(11): 1380–9.

BALSAC. 2012. Rapport annuel 2011-2012. Available at : http://balsac.uqac.ca/bibliographie/ publications-internes/ [Accessed: 2 Mar 2013].

BALSAC. 2014a. BALSAC Population database Available at: http://balsac.uqac.ca/english/

BALSAC. 2014b. BALSAC Population databse Available at: http://balsac.uqac.ca [Accessed: 3 Apr 2014].

BALSAC. 2014c. Rapport annuel 2013-2014. Available at : http://balsac.uqac.ca/bibliographie/publications-internes/ [Accessed: 1 Apr 2015].

Beaulieu CL, Majewski J, Schwartzentruber J, Samuels ME, Fernandez BA, Bernier FP, Brudno M, Knoppers B, Marcadier J, Dyment D, *et al.* 2014. FORGE Canada Consortium: Outcomes of a 2-Year National Rare-Disease Gene-Discovery Project. *Am J Hum Genet* 94(6): 809–17.

Bergeron J, Vézina H, Houde L, and Tremblay M. 2008. La contribution des Acadiens au peuplement des régions du Québec. *Cah québécois démographie* 37(1): 181–204.

Bernard G, Thiffault I, Tetreault M, Putorti ML, Bouchard I, Sylvain M, Melançon S, Laframboise R, Langevin P, Bouchard J-P, *et al.* 2010. Tremor-ataxia with central hypomyelination (TACH) leukodystrophy maps to chromosome 10q22.3-10q23.31. *Neurogenetics* 11(4): 457–64.

Bherer C, Labuda D, Roy-Gagnon M-H, Houde L, Tremblay M, and Vézina H. 2011. Admixed ancestry and stratification of Quebec regional populations. *Am J Phys Anthropol* 144(3): 432–41.

Bittles AH and Black ML. 2010. Consanguinity, human evolution and complex diseases. *Proc Natl Acad Sci U S A* 107(Suppl 1): 1779–86.

Black FL and Hedrick PW. 1997. Strong balancing selection at HLA loci: evidence from segregation in South Amerindian families. *Proc Natl Acad Sci U S A* 94(23): 12452–6.

Blanchette R. 2009. L'Outaouais. Québec: Les Presses de l'Université Laval.

Bouchard G. 2004. Genetic information and the risk of collective stigmatization. The case of the Saguenay-Lake St. John region (Quebec). *Med Sci (Paris)* 20(10): 933–4.

Bouchard JP, Barbeau A, Bouchard R, Paquet M, and Bouchard RW. 1979. A cluster of Friedreich's ataxia in Rimouski, Québec. *Can J Neurol Sci* 6(2): 205–8.

Bouchard J, Bedard P, and Bouchard R. 1980. Study of a family with progressive ataxia, tremor and severe distal amyotrophy. *Can J Neurol Sci* 7(4): 345–9.

Bouchard G and Braekeleer M De. 1990. Homogénéité ou diversité? L'histoire de la population du Québec revue à travers ses gènes. *Soc Hist Soc* 23(46): 325–61.

Bouchard G and Braekeleer M De. 1991. Histoire d'un génome : population et génétique dans l'est du Québec. Sillery: Presses de l'Université du Québec.

Bouchard G, Laberge C, and Scriver CR. 1988. Reproduction démographique et transmission génétique dans le Nord-Est de la province de Québec (18e-20e siècles). *Eur J Popul* 4(1): 39–67.

Bouchard G, Roy R, Casgrain B, and Hubert M. 1989. Fichiers de population et structure de gestion de bases de données : le fichier-réseau BALSAC et le système INGRES/INGRID. *Hist Mes* 4(1-2): 39–57.

Bourgain C and Génin E. 2005. Complex trait mapping in isolated populations: Are specific statistical methods required? *Eur J Hum Genet* 13(6): 698–706.

Braekeleer M De. 1991. Hereditary disorders in Saguenay-Lac-St-Jean (Quebec, Canada). *Hum Hered* 41(3): 141–6.

Braekeleer M De, Dallaire A, and Mathieu J. 1993a. Genetic epidemiology of sensorimotor polyneuropathy with or without agenesis of the corpus callosum in northeastern Quebec. *Hum Genet* 91(3): 223–7.

Braekeleer M De and Gauthier S. 1996. Autosomal recessive disorders in Saguenay-Lac-Saint-Jean (Quebec, Canada): a study of inbreeding. *Ann Hum Genet* 60(Pt 1): 51–6.

Braekeleer M De, Giasson F, Mathieu J, Roy M, Bouchard JP, and Morgan K. 1993b. Genetic epidemiology of autosomal recessive spastic ataxia of Charlevoix-Saguenay in northeastern Quebec. *Genet Epidemiol* 10(1): 17–25.

Braekeleer M De, Hechtman P, Andermann E, and Kaplan F. 1992. The French Canadian Tay-Sachs disease deletion mutation: Identification of probable founders. *Hum Genet* 89(1): 83–7.

Braekeleer M De and Larochelle J. 1990. Genetic epidemiology of hereditary tyrosinemia in Quebec and in Saguenay-Lac-St-Jean. *Am J Hum Genet* 47(2): 302–7.

Braekeleer M de, Vigneault A, and Simard H. 1992. Population genetics of hereditary hemochromatosis in Saguenay Lac-Saint-Jean (Quebec, Canada). *Ann Genet* 35(4): 202–7.

Brais B, Xie YG, Sanson M, Morgan K, Weissenbach J, Korczyn AD, Blumen SC, Fardeau M, Tomé FM, and Bouchard JP. 1995. The oculopharyngeal muscular dystrophy locus maps to the region of the cardiac alpha and beta myosin heavy chain genes on chromosome 14q11.2-q13. *Hum Mol Genet* 4(3): 429–34.

Brewster DH, Fordyce A, Black RJ, Bradshaw N, Campbell J, Cetnarskyj R, Davidson R, Drummond S, Garcia S, Gibbons B, *et al.* 2004. Impact of a cancer registry-based genealogy service to support clinical genetics services. *Fam Cancer* 3(2): 139–41.

Brown MD, Glazner CG, Zheng C, and Thompson EA. 2012. Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics* 190(4): 1447–60.

Browning SR and Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81(5): 1084–97.

Browning SR and Browning BL. 2010. High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet* 86(4): 526–39.

Browning SR and Browning BL. 2011a. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* 12(10): 703–14.

Browning BL and Browning SR. 2011b. A fast, powerful method for detecting identity by descent. *Am J Hum Genet* 88(2): 173–82.

Browning SR and Browning BL. 2012. Identity by descent between distant relatives: detection and applications. *Annu Rev Genet* 46: 617–33.

Browning BL and Browning SR. 2013a. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* 194(2): 459–71.

Browning BL and Browning SR. 2013b. Detecting identity by descent and estimating genotype error rates in sequence data. *Am J Hum Genet* 93(5): 840–51.

Browning SR and Thompson EA. 2012. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics* 190(4): 1521–31.

Bygren L, Tinghög P, Carstensen J, Edvinsson S, Kaati G, Pembrey ME, and Sjöström M. 2014. Change in paternal grandmothers´ early food supply influenced cardiovascular mortality of the female grandchildren. *BMC Genet* 15(1): 12.

Calò C, Melis A, Vona G, and Piras I. 2008. Review Synthetic Article: Sardinian Population (Italy): a Genetic Review. *Int J Mod Anthropol* 1(1): 39–64.

Camacho JA, Obie C, Biery B, Goodman BK, Hu CA, Almashanu S, Steel G, Casey R, Lambert M, Mitchell GA, *et al.* 1999. Hyperornithinaemia-hyperammonaemia-homocitrullinuria syndrome is caused by mutations in a gene encoding a mitochondrial ornithine transporter. *Nat Genet* 22(2): 151–8.

Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, Barac L, Pericic M, Janicijevic B, Smolej-Narancic N, Polasek O, *et al.* 2007. Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. *Hum Mol Genet* 16: 233–41.

Canadian Partnership Against Cancer Corporation. 2015. Canadian Partnership Against Cancer Available at: http://www.partnershipagainstcancer.ca/ [Accessed: 2 Nov 2014].

Cannon Albright LA. 2006. Computerized genealogies linked to medical histories for research and clinical care--a national view. *AMIA Annu Symp Proc*: 1161–2.

Cannon Albright LA. 2008. Utah family-based analysis: Past, present and future. *Hum Hered* 65(4): 209–20.

Cannon-Albright LA, Dintelman S, Maness T, Backus S, Thomas A, and Meyer LJ. 2013. Creation of a national resource with linked genealogy and phenotypic data: the Veterans Genealogy Project. *Genet Med* 15(7): 541–7.

Carmi S, Wilton PR, Wakeley J, and Pe'er I. 2014. A renewal theory approach to IBD sharing. *Theor Popul Biol* 97: 35–48.

CARTaGENE. 2015. CARTaGENE Available at: www.cartagene.qc.ca [Accessed: 1 Jan 2015].

Carter KC, Byck S, Waters PJ, Richards B, Nowacki PM, Laframboise R, Lambert M, Treacy E, and Scriver CR. 1998. Mutation at the phenylalanine hydroxylase gene (PAH) and its use to document population genetic variation: the Quebec experience. *Eur J Hum Genet* 6(1): 61–70.

Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, Maillard T de, Grenier J-C, Gbeha E, Hamdan FF, Girard S, *et al.* 2013. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet* 9(9): e1003815.

Cavallone L, Arcand SL, Maugard CM, Nolet S, Gaboury LA, Mes-Masson AM, Ghadirian P, Provencher D, and Tonin PN. 2010. Comprehensive BRCA1 and BRCA2 mutation analyses and review of French Canadian families with at least three cases of breast cancer. *Fam Cancer* 9(4): 507–17.

Cazes P and Cazes M-H. 1996. Comment mesurer la profondeur généalogique d'une ascendance? *Popul (Fr Ed)* 51(1): 117–40.

Chappuis PO, Hamel N, Paradis AJ, Deschênes J, Robidoux A, Potvin C, Cantin J, Tonin P, Ghadirian P, and Foulkes WD. 2001. Prevalence of founder BRCA1 and BRCA2 mutations in unselected French Canadian women with breast cancer. *Clin Genet* 59(6): 418–23.

Charbonneau H. 1973. La population du Québec ; études rétrospectives. Montréal: Éditions du Boréal Express.

Charbonneau H, Desjardins B, Légaré J, and Denis H. 2000. The population of the St-Lawrence Valley, 1608-1760. In *A population history of North America* (eds. Haines MR, Steckel RH), pp.99–142. New York: Cambridge University Press.

Charlesworth D and Willis JH. 2009. The genetics of inbreeding depression. *Nat Rev Genet* 10(11): 783–96.

Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, Li L, Lancaster G, Yang X, Williams A, *et al.* 2011. China Kadoorie Biobank of 0.5 million people: Survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* 40(6): 1652–66.

Chetaille P, Preuss C, Burkhard S, Côté J-M, Houde C, Castilloux J, Piché J, Gosset N, Leclerc S, Wünnemann F, *et al.* 2014. Mutations in SGOL1 cause a novel cohesinopathy affecting heart and gut rhythm. *Nat Genet* 46(11): 1245–9.

Ciullo M, Bellenguez C, Colonna V, Nutile T, Calabria A, Pacente R, Iovino G, Trimarco B, Bourgain C, and Persico MG. 2006. New susceptibility locus for hypertension on chromosome 8q by efficient pedigree-breaking in an Italian isolate. *Hum Mol Genet* 15(10): 1735–43.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, and Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 15(11): 1496–502.

Clarke CA, Edwards JW, Haddock DRW, Howel-Evans AW, McConnell RB, and Sheppard PM. 1956. ABO Blood groups and secretor character in duodenal ulcer. *Br Med J* 2(4995): 725–31.

Clouston HR. 1929. A Hereditary Ectodermal Dystrophy. *Can Med Assoc J* 21(1): 18–31.

Cohen JC, Boerwinkle E, Mosley TH, and Hobbs HH. 2006. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *N Engl J Med* 354(12): 1264–72.

Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, and Hobbs HH. 2005. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* 37(2): 161–5.

Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella K V, *et al.* 2011. Variation in genome-wide mutation rates within and between human families. *Nat Genet* 43(7): 712–4.

Cordell HJ and Clayton DG. 2005. Genetic association studies. *Lancet* 366(9491): 1121–31.

Cruz AL. 2013. An Overview of Genetic Counseling in Cuba. *J Genet Couns* 22(6): 849–53.

Cunnington MS, Koref MS, Mayosi BM, Burn J, and Keavney B. 2010. Chromosome 9p21 SNPs associated with multiple disease phenotypes correlate with ANRIL expression.

PLoS Genet 6(4): e1000899.

Dadu RT and Ballantyne CM. 2014. Lipid lowering with PCSK9 inhibitors. *Nat Rev Cardiol* 11(10): 563–75.

Daigneault J, Aubin G, Simard F, and Braekeleer M De. 1991. Genetic epidemiology of cystic fibrosis in Saguenay-Lac-St-Jean (Quebec, Canada). *Clin Genet* 40(4): 298–303.

Daoust-Boisvert A. 2014. Étude - Des centaines de Québécois malades sans le savoir. *Le Devoir*: 23 avril 2014, A5.

Dawber TR, Meadors GF, and Moore FE. 1951. Epidemiological approaches to heart disease: the Framingham Study. *Am J public Heal Nations Heal* 41(3): 279–81.

Dawn Teare M and Barrett JH. 2005. Genetic linkage studies. *Lancet* 366(9490): 1036–44.

Daxinger L and Whitelaw E. 2010. Transgenerational epigenetic inheritance: More questions than answers. *Genome Res* 20(12): 1623–8.

Debray F-G, Lambert M, Lemieux B, Soucy JF, Drouin R, Fenyves D, Dubé J, Maranda B, Laframboise R, and Mitchell GA. 2008. Phenotypic variability among patients with hyperornithinaemia-hyperammonaemia-homocitrullinuria syndrome homozygous for the delF188 mutation in SLC25A15. *J Med Genet* 45(11): 759–64.

deCODE Genetics Inc. 2015. deCODE genetics Available at: www.decode.com [Accessed: 1 Jun 2015].

Delaneau O, Marchini J, and Zagury J-F. 2012. A linear complexity phasing method for thousands of genomes. *Nat Methods* 9(2): 179–81.

Desjardins B. 1998. Le Registre de la population du Québec ancien. *Ann Demogr Hist (Paris)* 2: 215–26.

Desjardins M, Bélanger J, Yves F, and Bernard H. 1999. Histoire de la Gaspésie. Sainte-Foy, Québec: Institut québécois de recherche sur la culture / Presses de l'Université Laval.

Diamond JM and Rotter JI. 1987. Observing the founder effect in human evolution. *Nature* 329(6135): 105–6.

Dombrowski C, Lévesque S, Morel ML, Rouillard P, Morgan K, and Rousseau F. 2002. Premutation and intermediate-size FMR1 alleles in 10572 males from the general population: loss of an AGG interruption is a late event in the generation of fragile X syndrome alleles. *Hum Mol Genet* 11(4): 371–8.

Donnelly KP. 1983. The probability that related individuals share some section of genome identical by descent. *Theor Popul Biol* 23(1): 34–63.

Donnelly P. 2008. Progress and challenges in genome-wide association studies in humans. *Nature* 456(7223): 728–31.

Dupré N, Bouchard J-P, Brais B, and Rouleau GA. 2006. Hereditary ataxia, spastic paraparesis and neuropathy in the French-Canadian population. *Can J Neurol Sci* 33(2): 149–57.

Dupré N, Bouchard JP, Cossette L, Brunet D, Vanasse M, Lemieux B, Mathon G, and Puymirat J. 1999. Clinical and electrophysiological study in French-Canadian population

with Charcot-Marie-tooth disease type 1A associated with 17p11.2 duplication. *Can J Neurol Sci* 26(3): 196–200.

Dupré N, Cossette L, Hand CK, Bouchard JP, Rouleau GA, and Puymirat J. 2001. A founder mutation in French-Canadian families with X-linked hereditary neuropathy. *Can J Neurol Sci* 28(1): 51–5.

Dupré N, Gros-Louis F, Chrestian N, Verreault S, Brunet D, Verteuil D De, Brais B, Bouchard JP, and Rouleau GA. 2007. Clinical and genetic study of autosomal recessive cerebellar ataxia type 1. *Ann Neurol* 62(1): 93–8.

Dupré N, Howard HC, Mathieu J, Karpati G, Vanasse M, Bouchard JP, Carpenter S, and Rouleau GA. 2003. Hereditary motor and sensory neuropathy with agenesis of the corpus callosum. *Ann Neurol* 54(1): 9–18.

Duquette P and Giard N. 1997. Hereditary ptosis of late onset: Early observations on oculopharyngeal muscular dystrophy in Quebec by Roma Amyot. *Neuromuscul Disord* 7(SUPPL. 1): 97–9.

Duquette A, Roddier K, McNabb-Baltar J, Gosselin I, St-Denis A, Dicaire M-J, Loisel L, Labuda D, Marchand L, Mathieu J, *et al.* 2005. Mutations in senataxin responsible for Quebec cluster of ataxia with neuropathy. *Ann Neurol* 57(3): 408–14.

Dyke B. 1999. PEDSYS: a pedigree data management system User's Manual. San Antonio: Texas Southwest Foundation for Biomedical Research, Population Genetics Laboratory Technical Report No. 2.

Engert JC, Bérubé P, Mercier J, Doré C, Lepage P, Ge B, Bouchard JP, Mathieu J, Melançon SB, Schalling M, *et al.* 2000. ARSACS, a spastic ataxia common in northeastern Québec, is caused by mutations in a new gene encoding an 11.5-kb ORF. *Nat Genet* 24(2): 120–5.

Fareed M and Afzal M. 2014. Evidence of inbreeding depression on height, weight, and body mass index: A population-based child cohort study. *Am J Hum Biol* 26(6): 784–95.

Farrer LA, Bowirrat A, Friedland RP, Waraska K, Korczyn AD, and Baldwin CT. 2003. Identification of multiple loci for Alzheimer disease in a consanguineous Israeli-Arab community. *Hum Mol Genet* 12(4): 415–22.

Fave MJ, Hodgkinson AJ, Goulet J, Grenier J, Gauvin H, Bruat V, Maillard T, Gbeha E, Hip-Ki E, Idhagdour Y, *et al.* 2015. High-coverage RNA sequencing reveals substantial variation associated with geography, environment, and endophenotypic variation. (Program #358). Presented at the *65th Annual Meeting of The American Society of Human Genetics, October 10, 2015 in Baltimore, MD, USA*.

Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinburgh* 52: 399–433.

Fougères D. 2012. Histoire de Montréal et de sa région. Québec: Les Presses de l'Université Laval.

Foulkes AS. 2009. Applied Statistical Genetics with R. New York, NY: Springer New York.

Frenette P. 1996. Histoire de la Côte-Nord. Sainte-Foy, Québec: Institut québécois de recherche sur la culture.

Gagne C, Brun LD, Julien P, Moorjani S, and Lupien PJ. 1989. Primary lipoprotein-lipase-activity deficiency: Clinical investigation of a French Canadian population. *Cmaj* 140(4): 405–11.

Gagnon A and Heyer E. 2001. Fragmentation of the Québec population genetic pool (Canada): evidence from the genetic contribution of founders per region in the 17th and 18th centuries. *Am J Phys Anthropol* 114(1): 30–41.

Gauvin H, Moreau C, Lefebvre J-F, Laprise C, Vézina H, Labuda D, and Roy-Gagnon M-H. 2014. Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. *Eur J Hum Genet* 22(6): 814–21.

Gauvreau D, Guérin M, and Hamel M. 1991. De Charlevoix au Saguenay: mesure et caractréristiques du mouvement migratoire avant 1911. In *Histoire d'un génome* (eds. Bouchard G, Braekeleer M De), pp.145–59. Sillery, Québec: Presses de l'Université du Québec.

Genovese G, Leibon G, Pollak MR, and Rockmore DN. 2010. Improved IBD detection using incomplete haplotype information. *BMC Genet* 11(1): 58.

Gibson J, Morton NE, and Collins A. 2006. Extended tracts of homozygosity in outbred human populations. *Hum Mol Genet* 15(5): 789–95.

Glazner C and Thompson EA. 2012. Improving pedigree-based linkage analysis by estimating coancestry among families. *Stat Appl Genet Mol Biol* 11(2).

Gluckman PD, Hanson MA, and Beedle AS. 2007. Early life events and their consequences for later disease: A life history and evolutionary perspective. *Am J Hum Biol* 19(1): 1–19.

Godard B, Marshall J, Laberge C, and Knoppers BM. 2004. Strategies for consulting with the community: the cases of four large-scale genetic databases. *Sci Eng Ethics* 10(3): 457–77.

Gouvernement du Québec. 2015. MSSS - Public health - Blood and Urine Newborn Screening Available at: http://www.msss.gouv.qc.ca/en/sujets/santepub/depistage-neonatal/sanguin-et-urinaire/ [Accessed: 1 Mar 2015].

Grace RJ. 2003. Irish immigration and settlement in a Catholic city: Quebec, 1842-61. *Can Hist Rev* 84(2): 217–51.

Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, Rodriguez-Flores JL, Kenny EE, Gignoux CR, Maples BK, Guiblet W, *et al.* 2013. Reconstructing Native American migrations from whole-genome and whole-exome data. *PLoS Genet* 9(12): e1004023.

Gros-Louis F, Dupré N, Dion P, Fox MA, Laurent S, Verreault S, Sanes JR, Bouchard J-P, and Rouleau GA. 2007. Mutations in SYNE1 lead to a newly discovered form of autosomal recessive cerebellar ataxia. *Nat Genet* 39(1): 80–5.

Grossniklaus U, Kelly WG, Kelly B, Ferguson-Smith AC, Pembrey M, and Lindquist S. 2013. Transgenerational epigenetic inheritance: how important is it? *Nat Rev Genet* 14(3): 228–35.

Gulcher JR, Kong A, and Stefansson K. 2001. The role of linkage studies for common diseases. *Curr Opin Genet Dev* 11(3): 264–7.

Gulcher J and Stefansson K. 1998. Population genomics: Laying the groundwork for genetic

disease modeling and targeting. *Clin Chem Lab Med* 36(8): 523–7.

Gusev A, Kenny EE, Lowe JK, Salit J, Saxena R, Kathiresan S, Altshuler DM, Friedman JM, Breslow JL, and Pe'er I. 2011. DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *Am J Hum Genet* 88(6): 706–17.

Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, and Pe'er I. 2008. Whole population, genome-wide mapping of hidden relatedness. *Genome Res* 19(2): 318–26.

Gusev A, Palamara PF, Aponte G, Zhuang Z, Darvasi A, Gregersen P, and Pe'er I. 2012. The architecture of long-range haplotypes shared within and across populations. *Mol Biol Evol* 29(2): 473–86.

Hahn AF, Brown WF, Koopman WJ, and Feasby TE. 1990. X-linked dominant hereditary motor and sensory neuropathy. *Brain* 113: 1511–25.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, and King MC. 1990. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 250(4988): 1684–9.

Han L and Abney M. 2011. Identity by descent estimation with dense genome-wide genotype data. *Genet Epidemiol* 35(6): 557–67.

Hannou SA, Wouters K, Paumelle R, and Staels B. 2015. Functional genomics of the CDKN2A/B locus in cardiovascular and metabolic disease: what have we learned from GWASs? *Trends Endocrinol Metab* 26(4): 176–84.

Hardy R and Séguin N. 2008. La Mauricie. Québec: Les Presses de l'Université Laval.

Harris K and Nielsen R. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet* 9(6): e1003521.

Hattersley AT and McCarthy MI. 2005. What makes a good genetic association study? *Lancet* 366(9493): 1315–23.

He D. 2013. IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. *Bioinformatics* 29(13): i162–70.

Henn BM, Hon L, Macpherson JM, Eriksson N, Saxonov S, Pe'er I, and Mountain JL. 2012. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One* 7(4): e34267.

Heutink P and Oostra BA. 2002. Gene finding in genetically isolated populations. *Hum Mol Genet* 11(20): 2507–15.

Heyer E. 1999. One founder/one gene hypothesis in a new expanding population: Saguenay (Quebec, Canada). *Hum Biol* 71(1): 99–109.

Heyer E, Tremblay M, and Desjardins B. 1997. Seventeenth-century European origins of hereditary diseases in the Saguenay population (Quebec, Canada). *Hum Biol* 69(2): 209–25.

Hill WG. 1993. Variation in genetic identity within kinships. *Heredity (Edinb)* 71(6): 652–3.

Hodgkinson A, Idaghdour Y, Gbeha E, Grenier J-C, Hip-Ki E, Bruat V, Goulet J-P, Malliard T de, and Awadalla P. 2014. High-resolution genomic analysis of human mitochondrial RNA sequence variation. *Science* 344(6182): 413–5.

Hoffman JI, Simpson F, David P, Rijks JM, Kuiken T, Thorne MAS, Lacy RC, and Dasmahapatra KK. 2014. High-throughput sequencing reveals inbreeding depression in a natural population. *Proc Natl Acad Sci U S A* 111(10): 3775–80.

Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, Zanon C, Magnusson OT, Helgason A, Saemundsdottir J, Gylfason A, *et al.* 2011. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* 43(4): 316–20.

Hood L and Flores M. 2012. A personal view on systems medicine and the emergence of proactive P4 medicine: Predictive, preventive, personalized and participatory. *N Biotechnol* 29(6): 613–24.

Hostetler JA. 1985. History and relevance of the Hutterite population for genetic studies. *Am J Med Genet* 22(3): 453–62.

Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, and Freimer NB. 1994. Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat Genet* 8(4): 380–6.

Howie BN, Donnelly P, and Marchini J. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies (NJ Schork, Ed). *PLoS Genet* 5(6): e1000529.

Howrigan DP, Simonson MA, and Keller MC. 2011. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. *BMC Genomics* 12(1): 460.

Huang DW, Sherman BT, and Lempicki RA. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1): 44–57.

Huff CD, Witherspoon DJ, Simonson TS, Xing J, Watkins WS, Zhang Y, Tuohy TM, Neklason DW, Burt RW, Guthery SL, *et al.* 2011. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res* 21(5): 768–74.

Hussin JG, Hodgkinson A, Idaghdour Y, Grenier J-C, Goulet J-P, Gbeha E, Hip-Ki E, and Awadalla P. 2015. Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat Genet* 47(4): 400–4.

Ilmonen P, Penn DJ, Damjanovich K, Clarke J, Lamborn D, Morrison L, Ghotbi L, and Potts WK. 2008. Experimental infection magnifies inbreeding depression in house mice. *J Evol Biol* 21(3): 834–41.

International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs R a, Belmont JW, Boudreau A, Hardenbol P, *et al.* 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164): 851–61.

International Human Genome Sequencing Consortium. 2004. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 431(7011): 931–45.

Ismail J, Jafar TH, Jafary FH, White F, Faruqui AM, and Chaturvedi N. 2004. Risk factors for

non-fatal myocardial infarction in young South Asian adults. *Heart* 90(3): 259–63.

Jacquard A. 1974. The Genetic Structure of Populations. Berlin: Springer-Verlag.

Jaquish CE. 2007. The Framingham Heart Study, on its way to becoming the gold standard for Cardiovascular Genetic Epidemiology? *BMC Med Genet* 8(1): 63.

Jenkins T. 1990. Medical genetics around the world Medical genetics in South Africa. *J Med Genet* 27(12): 760–79.

Jensen PB, Jensen LJ, and Brunak S. 2012. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 13(6): 395–405.

Jetté R, Gauvreau D, and Guérin M. 1991. Aux origines d'une région: le peuplement fondateur de Charlevoix avant 1850. In *Histoire d'un génome* (eds. Bouchard G, Braekeleer M De), pp.75–106. Sillery: Presses de l'Université du Québec.

Jirtle RL and Skinner MK. 2007. Environmental epigenomics and disease susceptibility. *Nat Rev Genet* 8(4): 253–62.

John SW, Rozen R, Scriver CR, Laframboise R, and Laberge C. 1990. Recurrent mutation, gene conversion, or recombination at the human phenylalanine hydroxylase locus: evidence in French-Canadians and a catalog of mutations. *Am J Hum Genet* 46(5): 970–4.

Jorde LB. 1982. The genetic structure of the Utah Mormons: migration analysis. *Hum Biol an Int Rec Res* 54(3): 583–97.

Joshi PK, Esko T, Mattsson H, Eklund N, Gandin I, Nutile T, Jackson AU, Schurmann C, Smith A V., Zhang W, *et al.* 2015. Directional dominance on stature and cognition in diverse human populations. *Nature* 523(7561): 459–62.

Joubert M, Eisenring JJ, Robb JP, and Andermann F. 1969. Familial agenesis of the cerebellar vermis. A syndrome of episodic hyperpnea, abnormal eye movements, ataxia, and retardation. *Neurology* 19(9): 813–25.

Karigl G. 1981. A recursive algorithm for the calculation of identity coefficients. *Ann Hum Genet* 45(Pt 3): 299–305.

Keats BJ, Elston RC, and Andermann E. 1987a. Pedigree discriminant analysis of two French Canadian Tay-Sachs families. *Genet Epidemiol* 4(2): 77–85.

Keats BJ, Ward LJ, Lu M, Krieger S, Wilensky MA, Forster-Gibson CJ, Roy M, Monté M, Barbeau A, and Simpson NE. 1987b. Linkage studies of Friedreich ataxia by means of blood-group and protein markers. *Am J Hum Genet* 41(4): 627–34.

Keller MC, Simonson MA, Ripke S, Neale BM, Gejman P V, Howrigan DP, Lee SH, Lencz T, Levinson DF, and Sullivan PF. 2012. Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. *PLoS Genet* 8(4): 1–11.

Keller MC, Visscher PM, and Goddard ME. 2011. Quantification of Inbreeding Due to Distant Ancestors and Its Detection Using Dense Single Nucleotide Polymorphism Data. *Genetics* 189(1): 237–49.

Kenny EE, Gusev A, Riegel K, Lütjohann D, Lowe JK, Salit J, Maller JB, Stoffel M, Daly MJ, Altshuler DM, *et al.* 2009. Systematic haplotype analysis resolves a complex plasma plant sterol locus on the Micronesian Island of Kosrae. *Proc Natl Acad Sci U S A*

106(33): 13886–91.

Kere J. 2001. Human population genetics: lessons from Finland. *Annu Rev Genomics Hum Genet* 2(1): 103–28.

Khoury MJ, Beaty TH, and Cohen BH. 1993. Fundamentals of genetic epidemiology. New York: Oxford University Press.

Khoury MJ, Cohen BH, Diamond EL, Chase GA, and McKusick VA. 1987. Inbreeding and prereproductive mortality in the Old Order Amish. I. Genealogic epidemiology of inbreeding. *Am J Epidemiol* 125(3): 453–61.

Kibar Z, Dubé MP, Powell J, McCuaïg C, Hayflick SJ, Zonana J, Hovnanian A, Radhakrishna U, Antonarakis SE, Benohanian A, *et al.* 2000. Clouston hidrotic ectodermal dysplasia (HED): genetic homogeneity, presence of a founder effect in the French Canadian population and fine genetic mapping. *Eur J Hum Genet* 8(5): 372–80.

Knoppers BM and Joly Y. 2007. Our social genome? *Trends Biotechnol* 25(7): 284–8.

Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, Olason PI, Ingason A, Steinberg S, Rafnar T, *et al.* 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 40(9): 1068–75.

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, *et al.* 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319): 1099–103.

Kouladjian K. 1986. Une mesure d'entropie généalogique. Chicoutimi: Programme de recherches en génétique humaine.

Kristensen TN, Sørensen P, Kruhøffer M, Pedersen KS, and Loeschcke V. 2005. Genome-wide analysis on inbreeding effects on gene expression in Drosophila melanogaster. *Genetics* 171(1): 157–67.

Kristensen TN, Sørensen P, Pedersen KS, Kruhøffer M, and Loeschcke V. 2006. Inbreeding by environmental interactions affect gene expression in Drosophila melanogaster. *Genetics* 173(3): 1329–36.

Kristiansson K, Naukkarinen J, and Peltonen L. 2008. Isolated populations and complex disease gene identification. *Genome Biol* 9(8): 109.

Ku CS, Naidoo N, Teo SM, and Pawitan Y. 2011. Regions of homozygosity and their impact on complex diseases and traits. *Hum Genet* 129(1): 1–15.

la Chapelle A de and Wright FA. 1998. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proc Natl Acad Sci U S A* 95(21): 12416–23.

Laberge C. 1969. Hereditary tyrosinemia in a French Canadian isolate. *Am J Hum Genet* 21(1): 36–45.

Laberge A-M, Jomphe M, Houde L, Vézina H, Tremblay M, Desjardins B, Labuda D, St-Hilaire M, Macmillan C, Shoubridge EA, *et al.* 2005a. A "Fille du Roy" introduced the T14484C Leber hereditary optic neuropathy mutation in French Canadians. *Am J Hum*

Genet 77(2): 313–7.

Laberge A-M, Michaud J, Richter A, Lemyre E, Lambert M, Brais B, and Mitchell G. 2005b. Population history and its impact on medical genetics in Quebec. *Clin Genet* 68(4): 287–301.

Labuda M, Fujiwara TM, Ross M V, Morgan K, Garcia-Heras J, Ledbetter DH, Hughes MR, and Glorieux FH. 1992. Two hereditary defects related to vitamin D metabolism map to the same region of human chromosome 12q13-14. *J Bone Miner Res* 7(12): 1447–53.

Lacroix L. 2001. Génétique communautaire: Une médecine plus individualisée. *La Presse*: 18 février 2014, C1.

LaFramboise T. 2009. Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Res* 37(13): 4181–93.

Lambert G, Sjouke B, Choque B, Kastelein JJP, and Hovingh GK. 2012. The PCSK9 decade: Thematic Review Series: New Lipid and Lipoprotein Targets for the Treatment of Cardiometabolic Diseases. *J Lipid Res* 53(12): 2515–24.

Lander ES. 1996. The new genomics: global views of biology. *Science* 274(5287): 536–9.

Lander ES and Botstein D. 1987. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science (80-)* 236(4808): 1567–70.

Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, and Sobel EM. 2013. Mendel: The Swiss army knife of genetic analysis programs. *Bioinformatics* 29(12): 1568–70.

Larmuseau MHD, Geystelen A Van, Oven M Van, and Decorte R. 2013a. Genetic genealogy comes of age: perspectives on the use of deep-rooted pedigrees in human population genetics. *Am J Phys Anthropol* 150(4): 505–11.

Larmuseau MHD, Vanoverbeke J, Geystelen A Van, Defraene G, Vanderheyden N, Matthys K, Wenseleers T, and Decorte R. 2013b. Low historical rates of cuckoldry in a Western European human population traced by Y-chromosome and genealogical data. *Proc R Soc B Biol Sci* 280(1772): 20132400.

Lavery J V, Grady C, Wahl ER, and Emanuel EJ. 2007. Ethical Issues in International Biomedical Research: A Casebook. New York: Oxford University Press.

Lawson DJ and Falush D. 2012. Population Identification Using Genetic Data. *Annu Rev Genomics Hum Genet* 13(1): 337–61.

Lawson DJ, Hellenthal G, Myers S, and Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet* 8(1): e1002453.

Lee W-J, Pollin TI, O'Connell JR, Agarwala R, and Schäffer AA. 2010. PedHunter 2.0 and its usage to characterize the founder structure of the Old Order Amish of Lancaster County. *BMC Med Genet* 11(1): 68.

Légaré J. 1988. A Population Register for Canada Under the French Regime: Context, Scope, Content and Applications. *Can Stud Popul* 15(1): 1–16.

Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, and Malhotra AK. 2007. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc Natl Acad Sci* 104(50): 19942–7.

Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, Hutnik K, Royrvik EC, Cunliffe B, Lawson DJ, *et al.* 2015. The fine-scale genetic structure of the British population. *Nature* 519(7543): 309–14.

Levesque S, Morin C, Guay S-P, Villeneuve J, Marquis P, Yik W, Jiralerspong S, Bouchard L, Steinberg S, Hacia JG, *et al.* 2012. A founder mutation in the PEX6 gene is responsible for increased incidence of Zellweger syndrome in a French Canadian population. *BMC Med Genet* 13(1): 72.

Li H, Glusman G, Hu H, Shankaracharya, Caballero J, Hubley R, Witherspoon D, Guthery SL, Mauldin DE, Jorde LB, *et al.* 2014. Relationship estimation from whole-genome sequence data. *PLoS Genet* 10(1): e1004144.

Li LH, Ho SF, Chen CH, Wei CY, Wong WC, Li LY, Hung SI, Chung WH, Pan WH, Lee MTM, *et al.* 2006. Long contiguous stretches of homozygosity in the human genome. *Hum Mutat* 27(11): 1115–21.

Li Y, Willer CJ, Ding J, Scheet P, and Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 34(8): 816–34.

Li Y, Willer C, Sanna S, and Abecasis G. 2009. Genotype Imputation. *Annu Rev Genomics Hum Genet* 10(1): 387–406.

Lilienfeld AM. 1961. Problems and areas in genetic-epidemiological field studies. *Ann N Y Acad Sci* 91: 797–805.

Linteau P-A. 2007. Brève histoire de Montréal. Montréal: Éditions du Boréal.

Lyons EJ, Frodsham AJ, Zhang L, Hill AVS, and Amos W. 2009. Consanguinity and susceptibility to infectious diseases in humans. *Biol Lett* 5(March): 574–6.

MacCluer JW, VandeBerg JL, Read B, and Ryder OA. 1986. Pedigree analysis by computer simulation. *Zoo Biol* 5(2): 147–60.

Macgregor S, Bellis C, Lea RA, Cox H, Dyer T, Blangero J, Visscher PM, and Griffiths LR. 2010. Legacy of mutiny on the Bounty: founder effect and admixture on Norfolk Island. *Eur J Hum Genet* 18(1): 67–72.

Macmillan C, Kirkham T, Fu K, Allison V, Andermann E, Chitayat D, Fortier D, Gans M, Hare H, Quercia N, *et al.* 1998. Pedigree analysis of French Canadian families with T14484C Leber's hereditary optic neuropathy. *Neurology* 50(2): 417–22.

Malécot G. 1948. Les mathématiques de l'hérédité. Paris: Masson.

Marchini J and Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11(7): 499–511.

Marchini J, Howie B, Myers S, McVean G, and Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39(7): 906–13.

Marin-Leblanc M, Perreault S, Bahroun I, Lapointe M, Mongrain I, Provost S, Turgeon J, Talajic M, Brugada R, Phillips M, *et al.* 2012. Validation of warfarin pharmacogenetic algorithms in clinical practice. *Pharmacogenomics* 13(1): 21–9.

Markus B, Birk OS, and Geiger D. 2011. Integration of SNP genotyping confidence scores in IBD inference. *Bioinformatics* 27(20): 2880–7.

Mascalzoni D, Janssens ACJW, Stewart A, Pramstaller P, Gyllensten U, Rudan I, Duijn CM van, Wilson JF, Campbell H, and Quillan RMC. 2010. Comparison of participant information and informed consent forms of five European studies in genetic isolated populations. *Eur J Hum Genet* 18(3): 296–302.

Masel J. 2011. Genetic drift. *Curr Biol* 21(20): R837–8.

Mathieu J and Prévost C. 2012. Epidemiological surveillance of myotonic dystrophy type 1: A 25-year population-based study. *Neuromuscul Disord* 22(11): 974–9.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, and Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9(5): 356–69.

McClellan J and King MC. 2010. Genetic heterogeneity in human disease. *Cell* 141(2): 210–7.

McGowan-Jordan J, Stoddard K, Podolsky L, Orrbine E, McLaine P, Town M, Goodyer P, MacKenzie a, and Heick H. 1999. Molecular analysis of cystinosis: probable Irish origin of the most common French Canadian mutation. *Eur J Hum Genet* 7(6): 671–8.

McKusick VA. 1978. Medical Genetic Studies of the Amish: Selected Papers. Baltimore: Johns Hopkins University Press.

MCKUSICK VA, HOSTETLER JA, and EGELAND JA. 1964. GENETIC STUDIES OF THE AMISH, BACKGROUND AND POTENTIALITIES. *Bull Johns Hopkins Hosp* 115: 203–22.

McKusick-Nathans Institute of Genetic Medicine. 2015. Online Mendelian Inheritance in Man, OMIM® Available at: http://omim.org/ [Accessed: 4 Jun 2015].

McQuillan R, Eklund N, Pirastu N, Kuningas M, McEvoy BP, Esko T, Corre T, Davies G, Kaakinen M, Lyytikäinen L-P, *et al.* 2012. Evidence of Inbreeding Depression on Human Height (G Gibson, Ed). *PLoS Genet* 8(7): e1002655.

McQuillan R, Leutenegger A, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, *et al.* 2008. Runs of homozygosity in European populations. *Am J Hum Genet* 83(3): 359–72.

Meijer IA, Cossette P, Roussel J, Benard M, Toupin S, and Rouleau GA. 2004. A novel locus for pure recessive hereditary spastic paraplegia maps to 10q22.1-10q24.1. *Ann Neurol* 56(4): 579–82.

Meijer IA, Dupré N, Brais B, Cossette P, St-Onge J, Rioux M-F, Benard M, and Rouleau GA. 2007. SPG4 founder effect in French Canadians with hereditary spastic paraplegia. *Can J Neurol Sci* 34(2): 211–4.

Meijer IA, Hand CK, Cossette P, Figlewicz DA, and Rouleau GA. 2002. Spectrum of SPG4 mutations in a large collection of North American families with hereditary spastic paraplegia. *Arch Neurol* 59(2): 281–6.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, and Ding W. 1994. A strong candidate for the breast and ovarian cancer

susceptibility gene BRCA1. *Science* 266(5182): 66–71.

Mitchell JJ, Capua a, Clow C, and Scriver CR. 1996. Twenty-year outcome analysis of genetic screening programs for Tay-Sachs and beta-thalassemia disease carriers in high schools. *Am J Hum Genet* 59(4): 793–8.

Moltke I, Albrechtsen A, Hansen TVO, Nielsen FC, and Nielsen R. 2011. A method for detecting IBD regions simultaneously in multiple individuals--with applications to disease genetics. *Genome Res* 21(7): 1168–80.

Montpetit A, Côté S, Brustein E, Drouin CA, Lapointe L, Boudreau M, Meloche C, Drouin R, Hudson TJ, Drapeau P, *et al.* 2008. Disruption of AP1S1, Causing a Novel Neurocutaneous Syndrome, Perturbs Development of the Skin and Spinal Cord (V van Heyningen, Ed). *PLoS Genet* 4(12): e1000296.

Moreau C, Bhérer C, Vézina H, Jomphe M, Labuda D, and Excoffier L. 2011. Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science (80-)* 334(6059): 1148–50.

Moreau C, Lefebvre J-F, Jomphe M, Bhérer C, Ruiz-Linares A, Vézina H, Roy-Gagnon M-H, and Labuda D. 2013. Native American admixture in the Quebec founder population. *PLoS One* 8(6): e65507.

Morgan K, Holmes TM, Schlaut J, Marchuk L, Kovithavongs T, Pazderka F, and Dossetor JB. 1980. Genetic variability of HLA in the Dariusleut Hutterites. A comparative genetic analysis of the Hutterities, the Amish, and other selected Caucasian populations. *Am J Hum Genet* 32(2): 246–57.

Morin C, Mitchell G, Larochelle J, Lambert M, Ogier H, Robinson BH, and Braekeleer M De. 1993. Clinical, metabolic, and genetic aspects of cytochrome C oxidase deficiency in Saguenay-Lac-Saint-Jean. *Am J Hum Genet* 53(2): 488–96.

Morton DH, Morton CS, Strauss KA, Robinson DL, Puffenberger EG, Hendrickson C, and Kelley RI. 2003. Pediatric medicine and the genetic disorders of the Amish and Mennonite people of Pennsylvania. *Am J Med Genet* 121C(1): 5–17.

Murray GGR, Woolhouse MEJ, Tapio M, Mbole-Kariuki MN, Sonstegard TS, Thumbi SM, Jennings AE, Wyk IC van, Chase-Topping M, Kiara H, *et al.* 2013. Genetic susceptibility to infectious disease in East African Shorthorn Zebu: a genome-wide analysis of the effect of heterozygosity and exotic introgression. *BMC Evol Biol* 13(1): 246.

Nalls MA, Guerreiro RJ, Simon-Sanchez J, Bras JT, Traynor BJ, Gibbs JR, Launer L, Hardy J, and Singleton AB. 2009. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics* 10(3): 183–90.

Neel J and Schull W. 1954. Human heredity. Chicago: University of Chicago Press.

Nevanlinna HR. 1972. The Finnish population structure. A genetic and genealogical study. *Hereditas* 71(2): 195–236.

Newman DL, Abney M, McPeek MS, Ober C, and Cox NJ. 2001. The importance of genealogy in determining genetic associations with complex traits. *Am J Hum Genet* 69(5): 1146–8.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M,

Bhattacharjee A, Eichler EE, *et al.* 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461(7261): 272–6.

O'Brien E, Jorde LB, Rönnlöf B, Fellman JO, and Eriksson AW. 1988. Founder effect and genetic disease in Sottunga, Finland. *Am J Phys Anthropol* 77(3): 335–46.

Orton NC, Innes a. M, Chudley AE, and Bech-Hansen NT. 2008. Unique disease heritage of the Dutch-German mennonite population. *Am J Med Genet Part A* 146(8): 1072–87.

Ostrer H. 2001. A genetic profile of contemporary Jewish populations. *Nat Rev Genet* 2(11): 891–8.

Palamara PF, Lencz T, Darvasi A, and Pe'er I. 2012. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet* 91(5): 809–22.

Palin K, Campbell H, Wright AF, Wilson JF, and Durbin R. 2011. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genet Epidemiol* 35(8): 853–60.

Palmer LJ and Cardon LR. 2005. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* 366(9492): 1223–34.

Panoutsopoulou K, Hatzikotoulas K, Xifara DK, Colonna V, Farmaki A, Ritchie GRS, Southam L, Gilly A, Tachmazidou I, Fatumo S, *et al.* 2014. Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat Commun* 5: 5345.

Pasmant E, Laurendeau I, Héron D, Vidaud M, Vidaud D, and Bièche I. 2007. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: Identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res* 67(8): 3963–9.

Patel RS, Asselbergs FW, Quyyumi AA, Palmer TM, Finan CI, Tragante V, Deanfield J, Hemingway H, Hingorani AD, and Holmes M V. 2014. Genetic variants at chromosome 9p21 and risk of first versus subsequent coronary heart disease events: A systematic review and meta-analysis. *J Am Coll Cardiol* 63(21): 2234–45.

Patterson N, Price AL, and Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* 2(12): e190.

Pelletier VA, Galéano N, Brochu P, Morin CL, Weber AM, and Roy CC. 1986. Secretory diarrhea with protein-losing enteropathy, enterocolitis cystica superficialis, intestinal lymphangiectasia, and congenital hepatic fibrosis: a new syndrome. *J Pediatr* 108(1): 61–5.

Peltonen L, Jalanko A and Varilo. 1999. Molecular genetics of the Finnish disease heritage. *Hum Mol Genet* 8(10): 1913–23.

Peltonen L, Palotie A, and Lange K. 2000. Use of population isolates for mapping complex traits. *Nat Rev Genet* 1(3): 182–90.

Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg N a, and Li JZ. 2012. Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* 91(2): 275–92.

Pemberton TJ, Wang C, Li JZ, and Rosenberg NA. 2010. Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *Am J Hum Genet* 87(4): 457–64.

Perola M, Sammalisto S, Hiekkalinna T, Martin NG, Visscher PM, Montgomery GW, Benyamin B, Harris JR, Boomsma D, Willemsen G, *et al.* 2007. Combined genome scans for body stature in 6,602 European twins: Evidence for common caucasian loci. *PLoS Genet* 3(6): 1019–28.

Petrucelli N, Daly M, and Feldman G. 2013. BRCA1 and BRCA2 hereditary breast and ovarian cancer. In *GeneReviews® [Internet]* (eds. Pagon RA, Adam MP, Ardinger HH, *et al.*), Seattle: University of Washington.

Phaneuf D, Lambert M, Laframboise R, Mitchell G, Lettre F, and Tanguay RM. 1992. Type 1 hereditary tyrosinemia. Evidence for molecular heterogeneity and identification of a causal mutation in a French Canadian patient. *J Clin Invest* 90(4): 1185–92.

Piché V and Bourdais C Le. 2003. La démographie québécoise : enjeux du XXIe siècle. Montréal: Presses de l'Université de Montréal.

Pilia G, Chen W-M, Scuteri A, Orrú M, Albai G, Dei M, Lai S, Usala G, Lai M, Loi P, *et al.* 2006. Heritability of Cardiovascular and Personality Traits in 6,148 Sardinians. *PLoS Genet* 2(8): e132.

Plante M, Claveau S, Lepage P, Lavoie E-M, Brunet S, Roquis D, Morin C, Vézina H, and Laprise C. 2008. Mucolipidosis II: a single causal mutation in the N-acetylglucosamine-1-phosphotransferase gene (GNPTAB) in a French Canadian founder population. *Clin Genet* 73(3): 236–44.

Pouliot S and Rousseau J. 2014. Rapport d'évaluation du Projet-pilote d'offre de tests de porteur pour quatre maladies héréditaires récessives au Saguenay–Lac-Saint-Jean (Institut national de santé publique du Québec, Ed). Montréal.

Pouyez C, Lavoie Y, and Bouchard G. 1983. Les Saguenayens : introduction à l'histoire des populations du Saguenay, XVIe-XXe siècles. Sillery, Québec: Presses de l'Université du Québec.

Powell JE, Visscher PM, and Goddard ME. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet* 11(11): 800–5.

PRDH-IGD. 2015. Research Program in Historical Demography Available at: http://www.genealogy.umontreal.ca/en/home [Accessed: 5 May 2015].

Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, and Stefansson K. 2011. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet* 7(2): e1001317.

Puffenberger EG. 2003. Genetic heritage of the Old Order Mennonites of southeastern Pennsylvania. *Am J Med Genet C Semin Med Genet* 121C(1): 18–31.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, Bakker PIW de, Daly MJ, *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3): 559–75.

Qian Y, Browning BL, and Browning SR. 2014. Efficient clustering of identity-by-descent between multiple individuals. *Bioinformatics* 30(7): 915–22.

QRS. 2015. Quebec Reference Sample: Population Genetics and Genetic Epidemiology in Quebec Available at: http://www.quebecgenpop.ca/ [Accessed: 10 May 2015].

R Core Team. 2015. R: A Language and Environment for Statistical Computing. Vienna, Austria. https://cran.r-project.org/.

Rahman P, Jones A, Curtis J, Bartlett S, Peddle L, Fernandez BA, and Freimer NB. 2003. The Newfoundland population: a unique resource for genetic investigation of complex diseases. *Hum Mol Genet* 12(suppl 2): R167–72.

Richler M, Milot J, Quigley M, and O'Regan S. 1991. Ocular manifestations of nephropathic cystinosis. The French-Canadian experience in a genetically homogeneous population. *Arch Ophthalmol* 109(3): 359–62.

Rivard SR, Lanzara C, Grimard D, Carella M, Simard H, Ficarella R, Simard R, D'Adamo AP, Férec C, Camaschella C, *et al.* 2003. Juvenile hemochromatosis locus maps to chromosome 1q in a French Canadian population. *Eur J Hum Genet* 11(8): 585–9.

Rivard SR, Mura C, Simard H, Simard R, Grimard D, Gac G Le, Raguenes O, Férec C, and Braekeleer M De. 2000. Mutation analysis in the HFE gene in patients with hereditary haemochromatosis in Saguenay-Lac-Saint-Jean (Quebec, Canada). *Br J Haematol* 108(4): 854–8.

Roberts DF. 1968. Genetic effects of population size reduction. *Nature* 220: 1084–8.

Roddier K, Thomas T, Marleau G, Gagnon AM, Dicaire MJ, St.-Denis A, Gosselin I, Sarrazin AM, Larbrisseau A, Lambert M, *et al.* 2005. Two mutations in the HSN2 gene explain the high prevalence of HSAN2 in French Canadians. *Neurology* 64(10): 1762–7.

Rodriguez JM, Bercovici S, Huang L, Frostig R, and Batzoglou S. 2015. Parente2: a fast and accurate method for detecting identity by descent. *Genome Res* 25(2): 280–9.

Rohde DLT, Olson S, and Chang JT. 2004. Modelling the recent common ancestry of all living humans. *Nature* 431(7008): 562–6.

Rosenblatt DS. 2013. Victor McKusick and the History of Medical Genetics. *J Med Genet* 50(9): 640.

Rousseau F, Rouillard P, Morel ML, Khandjian EW, and Morgan K. 1995. Prevalence of carriers of premutation-size alleles of the FMRI gene--and implications for the population genetics of the fragile X syndrome. *Am J Hum Genet* 57(5): 1006–18.

Roy-Gagnon M-H, Moreau C, Bherer C, St-Onge P, Sinnett D, Laprise C, Vézina H, and Labuda D. 2011. Genomic and genealogical investigation of the French Canadian founder population structure. *Hum Genet* 129(5): 521–31.

Roy-Gagnon M-H, Weir MR, Sorkin JD, Ryan KA, Sack PA, Hines S, Bielak LF, Peyser PA, Post W, Mitchell BD, *et al.* 2008. Genetic influences on blood pressure response to the cold pressor test: results from the Heredity and Phenotype Intervention Heart Study. *J Hypertens* 26(4): 729–36.

Rudan I. 2006. Health effects of human population isolation and admixture. *Croat Med J* 47(4): 526–31.

Rudan I, Biloglav Z, Vorko-Jović A, Kujundzić-Tiljak M, Stevanović R, Ropac D, Puntarić D,

Cucević B, Salzer B, and Campbell H. 2006. Effects of inbreeding, endogamy, genetic admixture, and outbreeding on human health: a (1001 Dalmatians) study. *Croat Med J* 47(4): 601–10.

Rudan I, Rudan D, Campbell H, Carothers A, Wright A, Smolej-Narancic N, Janicijevic B, Jin L, Chakraborty R, Deka R, *et al.* 2003a. Inbreeding and risk of late onset complex disease. *J Med Genet* 40(12): 925–32.

Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, Janicijevic B, and Rudan P. 2003b. Inbreeding and the genetic complexity of human hypertension. *Genetics* 163(3): 1011–21.

Ruderfer DM, Lim ET, Genovese G, Moran JL, Hultman CM, Sullivan PF, McCarroll SA, Holmans P, Sklar P, and Purcell SM. 2015. No evidence for rare recessive and compound heterozygous disruptive variants in schizophrenia. *Eur J Hum Genet* 23(4): 555–7.

Saba TG, Montpetit A, Verner A, Rioux P, Hudson TJ, Drouin R, and Drouin CA. 2005. An atypical form of erythrokeratodermia variabilis maps to chromosome 7q22. *Hum Genet* 116(3): 167–71.

Samani NJ and Schunkert H. 2008. Chromosome 9p21 and cardiovascular disease: the story unfolds. *Circ Cardiovasc Genet* 1(2): 81–4.

Scheet P and Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78(4): 629–44.

Schork NJ, Murray SS, Frazer KA, and Topol EJ. 2009. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19(3): 212–9.

Scriver CR. 1988. Human Genes: Determinants of Sick Populations and Sick Patients. *Can J Public Heal / Rev Can Santé Publique* 79(4): 222–4.

Scriver CR. 2001. H <scp>UMAN</scp> G <scp>ENETICS</scp> : Lessons from Quebec Populations [1]. *Annu Rev Genomics Hum Genet* 2(1): 69–101.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29(1): 308–11.

Shifman S and Darvasi A. 2001. The value of isolated populations. *Nat Genet* 28(4): 309–10.

Sillanpää MJ. 2011. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity (Edinb)* 106(4): 511–9.

Simard LR, Rochette C, Semionov A, Morgan K, and Vanasse M. 1997. SMN(T) and NAIP mutations in Canadian families with spinal muscular atrophy (SMA): Genotype/phenotype correlations with disease severity. *Am J Med Genet* 72(1): 51–8.

Skolnick M. 1980. The Utah genealogical data base: A resource for genetic epidemiology. In *Cancer incidence in defined populations* (eds. Cairns J, Lyon JL, Skolnick M), pp.285–97. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory.

Skre H. 1975. Friedreich's ataxia in Western Norway. *Clin Genet* 7(4): 287–98.

Slattery ML and Kerber RA. 1993. A Comprehensive Evaluation of Family History and Breast

Cancer Risk. *JAMA* 270(13): 1563–8.

Smith CAB. 1953. The Detection of Linkage in Human Genetics. *J R Stat Soc Ser B* 15(2): 153–92.

Speed D and Balding DJ. 2015. Relatedness in the post-genomic era: is it still useful? *Nat Rev Genet* 16(1): 33–44.

Srour M, Hamdan FF, Schwartzentruber J a., Patry L, Ospina LH, Shevell MI, Desilets V, Dobrzeniecka S, Mathonnet G, Lemyre E, *et al.* 2012a. Mutations in TMEM231 cause Joubert syndrome in French Canadians. *J Med Genet* 49(10): 636–41.

Srour M, Schwartzentruber J, Hamdan FF, Ospina LH, Patry L, Labuda D, Massicotte C, Dobrzeniecka S, Capo-Chichi J-M, Papillon-Cavanagh S, *et al.* 2012b. Mutations in C5ORF42 cause Joubert syndrome in the French Canadian population. *Am J Hum Genet* 90(4): 693–700.

Statistics Canada. 2012. Census Profile. Census 2011. *Stat Canada Cat* (no. 98-316-XWE).

Stefansdottir V, Arngrimsson R, and Jonsson JJ. 2013a. Iceland—Genetic Counseling Services. *J Genet Couns* 22(6): 907–10.

Stefansdottir V, Johannsson OT, Skirton H, Tryggvadottir L, Tulinius H, and Jonsson JJ. 2013b. The use of genealogy databases for risk assessment in genetic health service: A systematic review. *J Community Genet* 4(1): 1–7.

Stevens EL, Heckenberg G, Baugher JD, Roberson EDO, Downey TJ, and Pevsner J. 2012. Consanguinity in Centre d'Étude du Polymorphisme Humain (CEPH) pedigrees. *Eur J Hum Genet* 20(6): 657–67.

Stevens EL, Heckenberg G, Roberson EDO, Baugher JD, Downey TJ, and Pevsner J. 2011. Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genet* 7(9): e1002287.

Strassmann BI, Kurapati NT, Hug BF, Burke EE, Gillespie BW, Karafet TM, and Hammer MF. 2012. Religion as a means to assure paternity. *Proc Natl Acad Sci* 109(25): 9781–5.

Swede H, Stone CL, and Norwood AR. 2007. National population-based biobanks for genetic research. *Genet Med* 9(3): 141–9.

Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zöllner S, Rosenberg NA, and Li JZ. 2013. Long runs of homozygosity are enriched for deleterious variation. *Am J Hum Genet* 93(1): 90–102.

Tataru P, Nirody JA, and Song YS. 2014. diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics* 30(23): 3430–1.

Tewhey R, Bansal V, Torkamani A, Topol EJ, and Schork NJ. 2011. The importance of phase information for human genomics. *Nat Rev Genet* 12(3): 215–23.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437(7063): 1299–320.

Thomas DC. 2004. Statistical methods in genetic epidemiology. Oxford: Oxford University Press.

Thompson EA. 1986. Pedigree Analysis in Human Genetics. Baltimore: Johns Hopkins University Press.

Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, and Risch N. 2012. Estimating kinship in admixed populations. *Am J Hum Genet* 91(1): 122–38.

Thorsson B, Sigurdsson G, and Gudnason V. 2003. Systematic family screening for familial hypercholesterolemia in Iceland. *Arterioscler Thromb Vasc Biol* 23(2): 335–8.

Tonin PN, Mes-Masson AM, Futreal PA, Morgan K, Mahon M, Foulkes WD, Cole DE, Provencher D, Ghadirian P, and Narod SA. 1998. Founder BRCA1 and BRCA2 mutations in French Canadian breast and ovarian cancer families. *Am J Hum Genet* 63(5): 1341–51.

Tremblay M, Arsenault J, and Heyer É. 2003. Les probabilités de transmission des gènes fondateurs dans cinq populations régionales du Québec. *Population (Paris)* 58(3): 403–23.

Tremblay M, Bouhali T, Gaudet D, and Brisson D. 2014. Genealogical analysis as a new approach for the investigation of drug intolerance heritability. *Eur J Hum Genet* 22(7): 916–22.

Tremblay M, Letendre M, Houde L, and Vézina H. 2008. The Contribution of Irish Immigrants to the Quebec (Canada) Gene Pool: An Estimation Using Data from Deep-Rooted Genealogies. *Eur J Popul / Rev Eur Démographie* 25(2): 215–33.

Tulinius H. 2011. Multigenerational information: the example of the Icelandic Genealogy Database. In *Methods in Biobanking* (ed. Dillner J), pp.221–9. New York: Humana Press.

Uimari P, Kontkanen O, Visscher PM, Pirskanen M, Fuentes R, and Salonen JT. 2005. Genome-wide linkage disequilibrium from 100,000 SNPs in the East Finland founder population. *Twin Res Hum Genet* 8(3): 185–97.

Vacic V, Ozelius LJ, Clark LN, Bar-Shira A, Gana-Weisz M, Gurevich T, Gusev A, Kedmi M, Kenny EE, Liu X, *et al.* 2014. Genome-wide mapping of IBD segments in an Ashkenazi PD cohort identifies associated haplotypes. *Hum Mol Genet* 23(17): 4693–702.

Venken T and Del-Favero J. 2007. Chasing genes for mood disorders and schizophrenia in genetically isolated populations. *Hum Mutat* 28(12): 1156–70.

Verhave JC, Troyanov S, Mongeau F, Fradette L, Bouchard J, Awadalla P, and Madore F. 2014. Prevalence, Awareness, and Management of CKD and Cardiovascular Risk Factors in Publicly Funded Health Care. *Clin J Am Soc Nephrol* 9(4): 713–9.

Verweij KJH, Abdellaoui A, Veijola J, Sebert S, Koiranen M, Keller MC, Järvelin MR, and Zietsch BP. 2014. The association of genotype-based inbreeding coefficient with a range of physical and psychological human traits. *PLoS One* 9(7).

Vézina H, Durocher F, Dumont M, Houde L, Szabo C, Tranchant M, Chiquette J, Plante M, Laframboise R, Lépine J, *et al.* 2005a. Molecular and genealogical characterization of the R1443X BRCA1 mutation in high-risk French-Canadian breast/ovarian cancer families. *Hum Genet* 117(2-3): 119–32.

Vézina H, Heyer É, Fortier I, Ouellette G, Robitaille Y, and Gauvreau D. 1999. A

genealogical study of Alzheimer disease in the Saguenay region of Quebec. *Genet Epidemiol* 16(4): 412–25.

Vézina H, Jomphe M, Lavoie È-M, Moreau C, and Labuda D. 2012. L'apport des données génétiques à la mesure généalogique des origines amérindiennes des Canadiens français. *Cah québécois démographie* 41(1): 87–105.

Vézina H, Tremblay M, Desjardins B, and Houde L. 2005b. Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. *Cah québécois démographie* 34(2): 235–58.

Vézina H, Tremblay M, and Houde L. 2004. Mesures de l'apparentement biologique au Saguenay-Lac-St-Jean (Québec, Canada) à partir de reconstitutions généalogiques. *Ann Demogr Hist (Paris)* 2(108): 67–83.

Vézina H, Tremblay M, Lavoie È-M, and Labuda D. 2014. Concordance entre origine ethnique déclarée et origines ancestrales chez les Gaspésiens. *Population (Paris)* 69(1): 7–28.

Visscher PM, Brown MA, McCarthy MI, and Yang J. 2012. Five Years of GWAS Discovery. *Am J Hum Genet* 90(1): 7–24.

Visscher PM, Hill WG, and Wray NR. 2008. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* 9(4): 255–66.

Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, Montgomery GW, and Martin NG. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* 2(3): e41.

Vohl MC, Moorjani S, Roy M, Gaudet D, Torres AL, Minnich A, Gagné C, Tremblay G, Lambert M, Bergeron J, *et al.* 1997. Geographic distribution of French-Canadian low-density lipoprotein receptor gene mutations in the Province of Quebec. *Clin Genet* 52(1): 1–6.

Vuillaumier-Barrot S, Bizec C Le, Lonlay P de, Barnier A, Mitchell G, Pelletier V, Prevost C, Saudubray JM, Durand G, and Seta N. 2002. Protein losing enteropathy-hepatic fibrosis syndrome in Saguenay-Lac St-Jean, Quebec is a congenital disorder of glycosylation type Ib. *J Med Genet* 39(11): 849–51.

Wakeley J, Nielsen R, Liu-Cordero SN, and Ardlie K. 2001. The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. *Am J Hum Genet* 69(6): 1332–47.

Wang K, Li M, and Hakonarson H. 2010. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16): 1–7.

Weinreich M and Frishman WH. 2014. Antihyperlipidemic therapies targeting PCSK9. *Cardiol Rev* 22(3): 140–6.

Weir BS, Anderson AD, and Hepler AB. 2006. Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet* 7(10): 771–80.

Weischenfeldt J, Symmons O, Spitz F, and Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* 14(2): 125–38.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, *et al.* 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42(D1): 1001–6.

Westerlind H, Imrell K, Ramanujam R, Myhr K, Celius EG, Harbo HF, Oturai AB, Hamsten A, Alfredsson L, Olsson T, *et al.* 2015. Identity-by-descent mapping in a Scandinavian multiple sclerosis cohort. *Eur J Hum Genet* 23(5): 688–92.

Wigginton JE, Cutler DJ, and Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 76(5): 887–93.

Woodley MA. 2009. Inbreeding depression and IQ in a study of 72 countries. *Intelligence* 37(3): 268–76.

World Health Organization. 2015. WHO | Cardiovascular diseases (CVDs). Available at: http://www.who.int/mediacentre/factsheets/fs317/en/ [Accessed: 30 Mar 2015].

Wright S. 1922. Coefficients of Inbreeding and Relationship. *Am Nat* 56(645): 330–8.

Wright AF, Carothers AD, and Pirastu M. 1999. Population choice in mapping genes for complex diseases. *Nat Genet* 23(4): 397–404.

Yotova V, Labuda D, Zietkiewicz E, Gehl D, Lovell A, Lefebvre JF, Bourgeois S, Lemieux-Blanchard É, Labuda M, Vézina H, *et al.* 2005. Anatomy of a founder effect: Myotonic dystrophy in Northeastern Quebec. *Hum Genet* 117(2-3): 177–87.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, *et al.* 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38(2): 203–8.

Zuk O, Hechter E, Sunyaev SR, and Lander ES. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci* 109(4): 1193–8.

Appendix A. Ethical certificates

Le 14 février 2012

Madame Marie-Hélène Roy-Gagnon
Pavillon Vidéotron
Étage B Bloc 4

OBJET: Titre du projet: Vers une utilisation optimale des ressources généalogiques de la population québécoise en épidémiologie génétique

No. de dossier: 3455

Responsables du projet: Marie-Hélène Roy-Gagnon Ph. D., chercheuse principale. Collaborateurs: Hélène Vézina et Damian Labuda

CHU Sainte-Justine
Le centre hospitalier universitaire mère-enfant

Pour l'amour des enfants

Université de Montréal

Madame,

Votre projet cité en rubrique a été approuvé par le comité d'éthique de la recherche en date du 9 février 2012. Vous trouverez ci-joint la liste des documents approuvés. Nous vous demandons de nous fournir une lettre d'autorisation de BALSAC spécifiquement pour votre projet, et ce, avant de débuter l'utilisation des données.
Notez que pour une collaboration avec un (ou plusieurs) tiers (institutions ou entreprises privées) impliquant des transferts de fonds et/ou données et/ou matériel biologique, une entente (contrat) doit être conclue avec le Bureau des ententes de recherche (BER).

Tous les projets de recherche impliquant des sujets humains doivent être réexaminés annuellement et la durée de l'approbation de votre projet sera effective jusqu'au **9 février 2013**. Notez qu'il est de votre responsabilité de soumettre une demande au comité pour que votre projet soit renouvelé avant la date d'expiration mentionnée. Il est également de votre responsabilité d'aviser le comité dans les plus brefs délais de toute modification au projet ainsi que de tout effet secondaire survenu dans le cadre de la présente étude.

Nous vous souhaitons bonne chance dans la réalisation de votre projet et vous prions de recevoir nos meilleures salutations.

Jean-Marie Therrien, Ph.D., éthicien
Président du Comité d'éthique de la recherche

JMT/mhl
BER

3175, Côte-Sainte-Catherine
Montréal (Québec)
H3T 1C5

Liste des documents approuvés par le CÉR

CHU Sainte-Justine
Le centre hospitalier universitaire mère-enfant

Pour l'amour des enfants

Université de Montréal

Titre du projet:
Vers une utilisation optimale des ressources généalogiques de la population québécoise en épidémiologie génétique

No. de dossier: 3455

Date d'approbation : jeudi 09 février 2012

Responsables du projet: ROY-GAGNON MARIE-HÉLÈNE Ph. D., chercheuse principale. Collaborateurs: Hélène Vézina et Damian Labuda

Liste:

- Protocole de recherche non daté

3175, Côte-Sainte-Catherine
Montréal (Québec)
H3T 1C5

Le 19 juin 2012

Monsieur Philip Awadalla
Pavillon Vidéotron
Étage B Bloc 4

OBJET: Titre du projet: Caractérisation des contributions génomiques et environnementales à la variation des phénotypes cardiovasculaires et métaboliques dans la population du Québec
No. de dossier: 3406
Responsables du projet: Philip Awadalla Ph. D., chercheur principal

CHU Sainte-Justine
Le centre hospitalier universitaire mère-enfant

Pour l'amour des enfants

Université de Montréal

Monsieur,

Votre projet cité en rubrique a été approuvé par le comité d'éthique de la recherche en date du 18 juin 2012. Vous trouverez ci-joint la liste des documents approuvés. Notez que pour une collaboration avec un (ou plusieurs) tiers (institutions ou entreprises privées) impliquant des transferts de fonds et/ou données et/ou matériel biologique, une entente (contrat) doit être conclue avec le Bureau des ententes de recherche (BER).

Tous les projets de recherche impliquant des sujets humains doivent être réexaminés annuellement et la durée de l'approbation de votre projet sera effective jusqu'au **18 juin 2013**. Notez qu'il est de votre responsabilité de soumettre une demande au comité pour que votre projet soit renouvelé avant la date d'expiration mentionnée. Il est également de votre responsabilité d'aviser le comité dans les plus brefs délais de toute modification au projet ainsi que de tout effet secondaire survenu dans le cadre de la présente étude.

Nous vous souhaitons bonne chance dans la continuité de votre projet et vous prions de recevoir nos meilleures salutations.

Jean-Marie Therrien, Ph.D., éthicien
Président du Comité d'éthique de la recherche

JMT/jda
c.c.: BER

3175, Côte-Sainte-Catherine
Montréal (Québec)
H3T 1C5

Liste des documents approuvés par le CÉR

CHU Sainte-Justine
Le centre hospitalier universitaire mère-enfant

Pour l'amour des enfants

Université de Montréal

Titre du projet:
Caractérisation des contributions génomiques et environnementales à la variation des phénotypes cardiovasculaires et métaboliques dans la population du Québec

No. de dossier: 3406

Date d'approbation : lundi 18 juin 2012

Responsables du projet: AWADALLA PHILIP Ph. D., chercheur principal

Liste:

- Protocole de recherche non daté, approuvé le 18 juin 2012

- Brochure d'information pour les participants

3175; Côte-Sainte-Catherine
Montréal (Québec)
H3T 1C5